HOWREY SIMON ARNOLD & WHITE
HOWREY LLP
ATTORNEYS AT LAW

1299 PENNSYLVANIA AVE., NW
WASHINGTON, DC 20004-2402
PHONE 202.783.0800
FAX 202.383.6610
A LIMITED LIABILITY PARTNERSHIP

October 10, 2000

Commissioner for Patents
Washington, DC 20231

Box Patent Application

Re: U.S. Non-Provisional Utility Patent Application
Application No.: To Be Assigned
Filed: Herewith
Title: Annotated Plant Genes
Inventor: David K. KOVALIC *et al.*
Atty. Dkt. No.: 04983.0204.CPUS00/38-21(15097)D

Sir:

The following documents are forwarded herewith for appropriate action by the U.S. Patent and Trademark Office:

1. Utility Patent Application Transmittal (PTO/SB/05);

2. U.S. Utility Patent Application entitled:

   Annotated Plant Genes

   and naming as inventors:

   David K. KOVALIC and Jingdong LIU,

   the application consisting of:

   a. A specification containing:

      (i) 59,101 pages of a description prior to the claims;

      (ii) 2 pages of claims (10 claims); and

      (iii) a one (1) page abstract;

3. A CD-ROM containing the sequence listing;

4. Petition to Suspend Sequence Rules (in duplicate);

5. Our Check No. 337505 in the amount of $130.00 to cover the petition fee; and

6. Two (2) return postcards.

This application is being filed without an executed Declaration, and without payment of official fees.

To facilitate the processing of the accompanying application, the CD-ROM containing the sequence listing can be located in Box 1. In addition, again to facilitate processing of the accompanying application, the pages setting forth the claims can be found in Box 17.

The U.S. Patent and Trademark Office is hereby authorized to charge any fee deficiency, or credit any overpayment, to our Deposit Account No. 08-3038. A duplicate copy of this letter is enclosed.

It is respectfully requested that, of the two attached postcards, one be stamped with the filing date of these documents and returned to our courier, and the other, prepaid postcard, be stamped with the filing date and unofficial application number and returned as soon as possible.

Sincerely,

David R. Marsh (Reg. No. 41,408)
June E. Cohan (Reg. No. 43,741)

Enclosures

Please type a plus sign (+) inside this box → [+]

PTO/SB/05 (2/98)
Approved for use through 09/30/00. OMB 0651-0032
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

UTILITY PATENT APPLICATION TRANSMITTAL
(Only for new nonprovisional applications under 37 CFR 1.53(b))

Attorney Docket No.	04983.0204.CPUS01/38-21(15097)D
First Named Inventor or Application Identifier	KOVALIC
Title	Annotated Plant Genes
Express Mail Label No.	

APPLICATION ELEMENTS
See MPEP chapter 600 concerning utility patent application contents

ADDRESS TO: Assistant Commissioner for Patents
Box Patent Application
Washington, DC 20231

1. ☐ *Fee Transmittal Form (Form PTO-1082) *(Submit an original and a duplicate for fee processing)*
2. ☒ Specification [*Total Pages* 59104] *(preferred arrangement set forth below)*
   - Descriptive title of the Invention
   - Cross References to Related Applications
   - Statement Regarding Fed sponsored R&D
   - Reference to Microfiche Appendix
   - Background of the Invention
   - Brief Summary of the Invention
   - Brief Description of the Drawings (if filed)
   - Detailed Description
   - Claims
   - Abstract of the Disclosure
3. ☐ Drawing(s) *(35 USC 113)* [*Total Sheets*]
4. Oath or Declaration [*Total Pages*]
   a. ☐ Newly executed (original or copy)
   b. ☐ Copy from a prior application (37 CFR 1.63(d)) *(for continuation/divisional with Box 17 completed)* ***[Note Box 5 below]***
      i. ☐ DELETION OF INVENTOR(S) Signed statement attached deleting inventor(s) named in the prior application, see 37 CFR 1.63(d)(2) and 1.33(b).
5. ☐ Incorporation By Reference *(useable if Box 4b is checked)* The entire disclosure of the prior application, from which a copy of the oath or declaration is supplied under Box 4b, is considered as being part of the disclosure of the accompanying application and is hereby incorporated by reference therein.
6. ☐ Microfiche Computer Program *(Appendix)*
7. Nucleotide and/or Amino Acid Sequence Submission *(if applicable, all necessary)*
   a. ☒ Computer Readable Copy
   b. ☐ Paper Copy (identical to computer copy)
   c. ☐ Statement verifying identity of above copies

ACCOMPANYING APPLICATION PARTS

8. ☐ Assignment Papers (cover sheet & document(s))
9. ☐ 37 CFR 3.73(b) Statement *(when there is an assignee)* ☐ Power of Attorney
10. ☐ English Translation Document *(if applicable)*
11. ☐ Information Disclosure Statement (IDS)/PTO-1449 ☐ Copies of IDS Citations
12. ☐ Preliminary Amendment
13. ☒ Return Receipt Postcard (MPEP 503) (Two) *(should be specifically itemized)*
14. ☐ *Small Entity Statement(s) ☐ Statement filed in prior application, Status still proper and desired
15. ☐ Certified Copy of Priority Document(s) *(if foreign priority is claimed)*
16. ☒ Other: Petition to Suspend Sequence Rules; Check #337505 in the amount of $130.00

*NOTE FOR ITEMS 1 & 14: IN ORDER TO BE ENTITLED TO PAY SMALL ENTITY FEES, A SMALL ENTITY STATEMENT IS REQUIRED (37 C.F.R. § 1.27), EXCEPT IF ONE FILED IN A PRIOR APPLICATION IS RELIED UPON (37 C.F.R. § 1.28).

17. **If a CONTINUING APPLICATION,** *check appropriate box and supply the requisite information:*
☐ Continuation ☐ Divisional ☒ Continuation-in-part (CIP) of prior application No: 09/654,617 (unofficial serial number)
Prior Application Information: Examiner: N/A Group/Art Unit: N/A

18. CORRESPONDENCE ADDRESS

☒ *Customer Number or Bar Code Label* 22930 *(Insert Customer No. or Attach bar code label here)* *or* ☐ *Correspondence address below*

NAME	David R. Marsh HOWREY SIMON ARNOLD & WHITE, LLP				
ADDRESS	Box No. 34 1299 Pennsylvania Avenue, N.W.				
CITY	Washington	*STATE*	DC	*ZIP CODE*	20004-2402
COUNTRY	US	*TELEPHONE*	202-783-0800	*FAX*	202-383-7195

Name (Print/Type)	David R. Marsh/June E. Cohan	*Registration No. (Attorney/Agent)*	41,408/43,741
Signature		*Date*	*October 10, 2000*

Burden Hour Statement: This form is estimated to take 0.2 hours to complete. Time will vary depending upon the needs of the individual case. Any comments on the amount of time you are required to complete this form should be sent to the Chief Information Officer, Patent and Trademark Office, Washington, DC 20231. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Assistant Commissioner for Patents, Box Patent Application, Washington, DC 20231.

Please type a plus sign (+) inside this box → [+]

PTO/SB/05 (2/98)
Approved for use through 09/30/00. OMB 0651-0032
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE

JC406 U.S. PTO 09/684016 10/10/00

UTILITY PATENT APPLICATION TRANSMITTAL

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Attorney Docket No.	04983.0204.CPUS01/38-21(15097)D
First Named Inventor or Application Identifier	KOVALIC
Title	Annotated Plant Genes
Express Mail Label No.	

APPLICATION ELEMENTS

See MPEP chapter 600 concerning utility patent application contents

ADDRESS TO: Assistant Commissioner for Patents
Box Patent Application
Washington, DC 20231

1. ☐ *Fee Transmittal Form (Form PTO-1082)
   (Submit an original and a duplicate for fee processing)
2. ☒ Specification [*Total Pages* 59104]
   (preferred arrangement set forth below)
   - Descriptive title of the Invention
   - Cross References to Related Applications
   - Statement Regarding Fed sponsored R&D
   - Reference to Microfiche Appendix
   - Background of the Invention
   - Brief Summary of the Invention
   - Brief Description of the Drawings (if filed)
   - Detailed Description
   - Claims
   - Abstract of the Disclosure
3. ☐ Drawing(s) *(35 USC 113)* [*Total Sheets*]
4. ☐ Oath or Declaration [*Total Pages*]
   a. ☐ Newly executed (original or copy)
   b. ☐ Copy from a prior application (37 CFR 1.63(d))
   (for continuation/divisional with Box 17 completed)
   [Note Box 5 below]
   i. ☐ DELETION OF INVENTOR(S)
   Signed statement attached deleting inventor(s) named in the prior application, see 37 CFR 1.63(d)(2) and 1.33(b).
5. ☐ Incorporation By Reference *(useable if Box 4b is checked)*
   The entire disclosure of the prior application, from which a copy of the oath or declaration is supplied under Box 4b, is considered as being part of the disclosure of the accompanying application and is hereby incorporated by reference therein.
6. ☐ Microfiche Computer Program *(Appendix)*
7. Nucleotide and/or Amino Acid Sequence Submission
   (if applicable, all necessary)
   a. ☒ Computer Readable Copy
   b. ☐ Paper Copy (identical to computer copy)
   c. ☐ Statement verifying identity of above copies

ACCOMPANYING APPLICATION PARTS

8. ☐ Assignment Papers (cover sheet & document(s))
9. ☐ 37 CFR 3.73(b) Statement *(when there is an assignee)* ☐ Power of Attorney
10. ☐ English Translation Document *(if applicable)*
11. ☐ Information Disclosure Statement (IDS)/PTO-1449 ☐ Copies of IDS Citations
12. ☐ Preliminary Amendment
13. ☒ Return Receipt Postcard (MPEP 503) (Two)
    (should be specifically itemized)
14. ☐ *Small Entity Statement(s) ☐ Statement filed in prior application, Status still proper and desired
15. ☐ Certified Copy of Priority Document(s)
    (if foreign priority is claimed)
16. ☒ Other: Petition to Suspend Sequence Rules; Check #337505 in the amount of $130.00

*NOTE FOR ITEMS 1 & 14 IN ORDER TO BE ENTITLED TO PAY SMALL ENTITY FEES, A SMALL ENTITY STATEMENT IS REQUIRED (37 C.F.R. § 1.27), EXCEPT IF ONE FILED IN A PRIOR APPLICATION IS RELIED UPON (37 C.F.R. § 1.28).

17. **If a CONTINUING APPLICATION,** *check appropriate box and supply the requisite information:*
    ☐ Continuation ☐ Divisional ☒ Continuation-in-part (CIP) of prior application No: 09/654,617 (unofficial serial number)
    Prior Application Information: Examiner: N/A Group/Art Unit: N/A

18. CORRESPONDENCE ADDRESS

☒ *Customer Number or Bar Code Label* 22930 *(Insert Customer No. or Attach bar code label here)* *or* ☐ *Correspondence address below*

NAME	David R. Marsh				
	HOWREY SIMON ARNOLD & WHITE, LLP				
ADDRESS	Box No. 34				
	1299 Pennsylvania Avenue, N.W.				
CITY	Washington	*STATE*	DC	*ZIP CODE*	20004-2402
COUNTRY	US	*TELEPHONE*	202-783-0800	*FAX*	202-383-7195

Name (Print/Type)	David R. Marsh/June E. Cohan	*Registration No. (Attorney/Agent)*	41,408/43,741
Signature		*Date*	*October 10, 2000*

ANNOTATED PLANT GENES

CROSS REFERENCE TO RELATED APPLICATION

The present application is a continuation-in-part of Application No: 09/654,617 (unofficial serial number) (Attorney Docket No: 04983.0204.00US00/38-21(15097)D), entitled "Annotated Plant Genes," filed September 5, 2000, inventors David K. Kovalic and Jingdong Liu.

FIELD OF THE INVENTION

The present invention is in the field of plant biochemistry. More specifically the invention relates to nucleic acid sequences from plant cells, in particular, nucleic acid sequences from maize, teosinte, soybean, *Arabidopsis*, cotton, sorghum, rice and wheat. The invention encompasses nucleic acid molecules that encode proteins and fragments of proteins. In addition, the invention also encompasses proteins and fragments of proteins so encoded and antibodies capable of binding these proteins or fragments. The invention also relates to methods of using the nucleic acid molecules, proteins and fragments of proteins, and antibodies, for example for genome mapping, gene identification and analysis, plant breeding, preparation of constructs for use in plant gene expression, and transgenic plants.

BACKGROUND OF THE INVENTION

The identification and isolation of plant genes belonging to biochemical and regulatory pathways are important in the development of nutritionally and agriculturally enhanced crops and products. Such nucleic acid molecules can be used in a variety of applications. For example, a nucleic acid molecule or a collection of nucleic acid molecules may act as a marker for a developmental or commercially valuable trait such as disease resistance. Additionally, they may be used to obtain homologues in the same or a different species. Nucleic acid molecules comprising coding sequences may also aid in gene expression studies that allow the dissection and elucidation of commercially useful traits.

The present invention provides nucleic acid molecules that are drawn from maize, soybean, rice, cotton, sorghum, wheat *Arabidopsis* and teosinte. They exhibit significant homology with known nucleic acid sequences belonging to a variety of biochemical and regulatory pathways.

Descriptions of biochemical and regulatory pathways are available from a number of well known academic and research organizations. An exemplary listing of such pathways can be

found in US patent application Serial number 09/371,146, the entirety of which is herein incorporated herein by reference, see especially pages 1-312. Additionally, several web sites and databases contain information pertaining to biochemical pathways and regulatory pathways. Examples of such web sites or data bases include: http://cgsc.biology.yale.edu (the CGSC maintains a database of *E. coli* genetic information, including genotypes and reference information for the strains in the CGSC collection, gene names, properties, and linkage map, gene product information, and information on specific mutations); http://www.labmed.umn.edu (the University of Minnesota's Biocatalysis/Biodegradation web page provides a search engine for compounds, enzymes, microorganisms, chemical formulas CAS registry, EC accession and microbial biocatalytic reactions and biodegradation pathways primarily for xenobiotic, chemical compounds such methionine, and threonine); http://wit.mcs.anl.gov/WIT2 (this website provides a functional overview which outlines metabolic pathways for organisms such as *E. coli*); http://ecocyc.PangeaSystems.com/ecocyc/ecocyc.html (this web site provides an overview of an *E. coli* metabolic map); http://www.biology.UCSD.edu (this web site provides information on signal transduction in higher plants); http://geo.nihs.go.jp (the Japanese National Institute of Health Science server provides information particularly on cell signaling networks); http://gifts.univ-mrs.fr (the Gene Intereactions in Fly Trans-world Server provides information on gene interactions, mostly centered on *Drosophila* gene interactions); http://sdb.bio.purdue.edu (this web site provides a data base of Drosophila genes); http://genome-www.stanford.edu (Stanford Genomic Research web site provides information on for example, *Sacchromyces* and *Arabidopsis*); http://www.psynix.co.uk (this web site provides illustrations and computer models of various cytokinins); http://www.sdsc.edu/Kinases/pk_home.html (this web site provides information on the protein kinase family of enzymes); http://transfac.gbf-braunschweig.de (the GBF web site provides information on regulatory genomic signals and regions, in particular those that govern transcriptional control); http://www.gcrdb.uthscsa.edu (this web site provides information on G-protein coupled receptors); http://www.biochem.purdue.edu (this web site provides information on secondary metabolism in *Arabidopsis*); http://home.wxs.nl/~pvsanten/ mmp/mmp.html (this web site provides a flow chart of metabolic pathways); http://www.genome.ad.jp/kegg/regulation.html (this web site, the KEGG regulatory pathways web site, provides pathway maps, ortholog group tables, and molecular catalogs searchable data bases by enzyme, pathway, or EC

number); http://capsulapedia.uchicago.edu/ Capsulapedia/Metabolism/RegExpMet.shtml (this website provides expression information); http://www.zmbh.uni-heidelberg.de/M_pneumoniae/genome/META/ALL_META.GIF (this web site provides a graphic of metabolic pathways and the ways these pathways interact); http://moulon.inra.fr/cgi-bin/nph-acedb3.1/acedb/metabolisme (this web site provides information on *C. elegans* metabolic enzymes); http://www.gwu.edu/~mpb (this web site provides information on metabolic pathways); http://www.bic.nus.edu.sg/pathwaydb.html (this web site provides links to biological pathways, such as metabolic pathways, developmental pathways, signal-transduction pathways, and genetic regulatory circuits); and http://www.scri.sari.ac.uk/bpp/charttxt.htm (this web site provides graphics of the metabolic pathways of diseased potato).

SUMMARY OF THE INVENTION

The present invention provides a substantially purified nucleic acid molecule where the nucleic acid molecule comprises a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof or fragments of either.

The present invention provides a substantially purified first nucleic acid molecule, wherein the first nucleic acid molecule specifically hybridizes to a second nucleic acid molecule having a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof.

The present invention provides a marker nucleic acid molecule capable of detecting the level, pattern, occurrence or absence of a biochemical process, wherein the biochemical process is selected from the group consisting of photosynthetic activity, carbohydrate metabolism, amino acid synthesis or degradation, plant hormone or other regulatory molecules, phenolic metabolism, lipid metabolism, biosynthesis of tetrapyrroles, phytochrome metabolism, carbon assimilation, glycolysis metabolism, gluconeogenesis metabolism, sucrose metabolism, starch metabolism, phosphogluconate metabolism, galactomannan metabolism, raffinose metabolism, complex carbohydrate metabolism, phytic acid metabolism, methionine biosynthesis, methionine degradation, lysine metabolism, arginine metabolism, proline metabolism, glutamate/glutamine metabolism, aspartate/asparagine metabolism, cytokinin metabolism, gibberellin metabolism, ethylene metabolism, jasmonic acid metabolism, transcription factors, R-genes, plant proteases, protein kinases, antifungal proteins, nitrogen transporters, sugar transporters, shikimate metabolism, isoflavone

metabolism, phenylpropanoid metabolism, isoprenoid metabolism, alpha-oxidation lipid metabolism, fatty acid metabolism, glycolysis metabolism, gluconeogenesis metabolism, sucrose metabolism, sucrose catabolism, reductive pentose phosphate cycle, regulation of C3 photosynthesis, C4 pathway carbon assimilation, enzymes involved in the C4 pathway, carotenoid metabolism, tocopherol metabolism, phytosterol metabolism, brassinoid metabolism, and proline metabolism.

The present invention also provides a substantially purified protein or fragment thereof encoded by a first nucleic acid molecule which specifically hybridizes to a second nucleic acid molecule, the second nucleic acid molecule having a nucleic acid sequence selected from the group consisting of a complement of SEQ ID NO: 1 through SEQ ID NO:463,173.

The present invention also provides a substantially purified protein or fragment thereof encoded by a nucleic acid molecule comprising a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO:463,173.

The present invention also provides a purified antibody or fragment thereof which is capable of specifically binding to a protein or fragment thereof, wherein the protein or fragment thereof is encoded by a nucleic acid molecule comprising a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 463,173.

The present invention also provides a transformed plant having a nucleic acid molecule which comprises: (A) an exogenous promoter region which functions in a plant cell to cause the production of a mRNA molecule; (B) a structural nucleic acid molecule comprising a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 463,173; and (C) a 3’ non-translated sequence that functions in the plant cell to cause termination of transcription and addition of polyadenylated ribonucleotides to a 3’ end of the mRNA molecule.

The present invention also provides a transformed plant having a nucleic acid molecule which comprises: (A) an exogenous promoter region which functions in a plant cell to cause the production of a mRNA molecule; which is linked to (B) a nucleic acid molecule with a transcribed strand and a non-transcribed strand, wherein the transcribed strand is complementary to a nucleic acid molecule comprising a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 463,173 or fragment thereof; which is linked to (C) a 3' non-translated sequence that functions in plant cells to cause termination of transcription and addition of polyadenylated ribonucleotides to a 3' end of the mRNA molecule.

The present invention provides a microarray comprising a collection of nucleic acid molecules wherein the collection of nucleic acid molecules are capable of detecting or predicting a component or attribute of a biochemical process or activity, where the biochemical process or activity are selected from the group consisting of photosynthetic activity, carbohydrate metabolism, amino acid synthesis or degradation, plant hormone or other regulatory molecules, phenolic metabolism, lipid metabolism, biosynthesis of tetrapyrroles, phytochrome metabolism, carbon assimilation, glycolysis metabolism, gluconeogenesis metabolism, sucrose metabolism, starch metabolism, phosphogluconate metabolism, galactomannan metabolism, raffinose metabolism, complex carbohydrate metabolism, phytic acid metabolism, methionine biosynthesis, methionine degradation, lysine metabolism, arginine metabolism, proline metabolism, glutamate/glutamine metabolism, aspartate/asparagine metabolism, cytokinin metabolism, gibberellin metabolism, ethylene metabolism, jasmonic acid metabolism, transcription factors, R-genes, plant proteases, protein kinases, antifungal proteins, nitrogen transporters, sugar transporters, shikimate metabolism, isoflavone metabolism, phenylpropanoid metabolism, isoprenoid metabolism, α-oxidation lipid metabolism, fatty acid metabolism, glycolysis metabolism, gluconeogenesis metabolism, sucrose metabolism, sucrose catabolism, reductive pentose phosphate cycle, regulation of C3 photosynthesis, C4 pathway carbon assimilation, enzymes involved in the C4 pathway, carotenoid metabolism, tocopherol metabolism, phytosterol metabolism, brassinoid metabolism, and proline metabolism.

The present invention also provides a method for determining a level or pattern of a plant protein in a plant cell or plant tissue comprising: (A) incubating, under conditions permitting nucleic acid hybridization, a marker nucleic acid molecule having a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof or fragment of either, with a complementary nucleic acid molecule obtained from the plant cell or plant tissue, wherein nucleic acid hybridization between the marker nucleic acid molecule and the complementary nucleic acid molecule obtained from the plant cell or plant tissue permits the detection of the protein; (B) permitting hybridization between the marker nucleic acid molecule and the complementary nucleic acid molecule obtained from the plant cell or plant tissue; and (C) detecting the level or pattern of the complementary nucleic acid, wherein the detection of the complementary nucleic acid is predictive of the level or pattern of the protein.

The present invention also provides a method for determining a level or pattern of a protein in a plant cell or plant tissue under evaluation which comprises assaying the concentration of a molecule, whose concentration is dependent upon the expression of a gene, the gene specifically hybridizes to a nucleic acid molecule having a nucleic acid sequence selected from the group consisting of a complement of SEQ ID NO: 1 through SEQ ID NO: 463,173, in comparison to the concentration of that molecule present in a reference plant cell or a reference plant tissue with a known level or pattern of the protein, wherein the assayed concentration of the molecule is compared to the assayed concentration of the molecule in the reference plant cell or reference plant tissue with the known level or pattern of the protein.

The present invention provides a method of determining a mutation in a plant whose presence is predictive of a mutation affecting a level or pattern of a protein comprising the steps: (A) incubating, under conditions permitting nucleic acid hybridization, a marker nucleic acid selected from the group of marker nucleic acid molecules which specifically hybridize to a nucleic acid molecule having a nucleic acid sequence selected from the group of SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof and a complementary nucleic acid molecule obtained from the plant, wherein nucleic acid hybridization between the marker nucleic acid molecule and the complementary nucleic acid molecule obtained from the plant permits the detection of a polymorphism whose presence is predictive of a mutation affecting the level or pattern of the protein in the plant; (B) permitting hybridization between the marker nucleic acid molecule and the complementary nucleic acid molecule obtained from the plant; and (C) detecting the presence of the polymorphism, wherein the detection of the polymorphism is predictive of the mutation.

The present invention also provides a method of producing a plant containing an overexpressed protein comprising: (A) transforming the plant with a functional nucleic acid molecule, wherein the functional nucleic acid molecule comprises a promoter region, wherein the promoter region is linked to a structural region, wherein the structural region has a nucleic acid sequence selected from group consisting of SEQ ID NO: 1 through SEQ ID NO: 463,173; wherein the structural region is linked to a 3' non-translated sequence that functions in the plant to cause termination of transcription and addition of polyadenylated ribonucleotides to a 3' end of a mRNA molecule; and wherein the functional nucleic acid molecule results in overexpression of the protein; and (B) growing the transformed plant.

The present invention also provides a method of producing a plant containing reduced levels of a protein comprising: (A) transforming the plant with a functional nucleic acid molecule, wherein the functional nucleic acid molecule comprises a promoter region, wherein the promoter region is linked to a structural region, wherein the structural region comprises a nucleic acid molecule having a nucleic acid sequence selected from the group consisting of a complement of SEQ ID NO: 1 through SEQ ID NO: 463,173 or fragment thereof and the transcribed strand is complementary to an endogenous mRNA molecule; and wherein the transcribed nucleic acid molecule is linked to a 3' non-translated sequence that functions in the plant cell to cause termination of transcription and addition of polyadenylated ribonucleotides to a 3' end of a mRNA molecule; and (B) growing the transformed plant.

The present invention also provides a method of determining an association between a polymorphism and a plant trait comprising: (A) hybridizing a nucleic acid molecule specific for the polymorphism to genetic material of a plant, wherein the nucleic acid molecule has a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof or fragment of either; and (B) calculating the degree of association between the polymorphism and the plant trait.

The present invention also provides a method of isolating a nucleic acid comprising: (A) incubating under conditions permitting nucleic acid hybridization, a first nucleic acid molecule comprising a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof or fragment of either with a complementary second nucleic acid molecule obtained from a plant cell or plant tissue; (B) permitting hybridization between the first nucleic acid molecule and the second nucleic acid molecule obtained from the plant cell or plant tissue; and (C) isolating the second nucleic acid molecule.

The present invention also provides a method of analyzing the differences in the RNA profiles from more than one physiological source, the method comprising: a) obtaining a sample of ribonucleic acids from each of the physiological sources; b) generating a population of labeled nucleic acids for each of the physiological sources from said sample of ribonucleic acids; c) hybridizing the labeled nucleic acids for each of the physiological sources to an array of nucleic acid molecules stably associated with the surface of a substrate to produce a hybridization pattern for each of the physiological sources; said stably associated nucleic acid molecules selected from

the group consisting of SEQ ID NO: 1 through SEQ ID NO:463,173 or fragments thereof and d) comparing the hybridization patterns for each of the different physiological sources.

DETAILED DESCRIPTION OF THE INVENTION

One skilled in the art can refer to general reference texts for detailed descriptions of known techniques discussed herein or equivalent techniques. These texts include *Current Protocols in Molecular Biology* Ausubel, *et al.*, eds., John Wiley & Sons, N. Y. (1989), and supplements through September (1998), *Molecular Cloning, A Laboratory Manual*, Sambrook *et al*, 2nd Ed., Cold Spring Harbor Press, Cold Spring Harbor, New York (1989), *Genome Analysis: A Laboratory Manual 1: Analyzing DNA*, Birren *et al.*, Cold Spring Harbor Press, Cold Spring Harbor, New York (1997); *Genome Analysis: A Laboratory Manual 2: Detecting Genes*, Birren *et al.*, Cold Spring Harbor Press, Cold Spring Harbor, New York (1998); *Genome Analysis: A Laboratory Manual 3: Cloning Systems*, Birren *et al.*, Cold Spring Harbor Press, Cold Spring Harbor, New York (1999); *Genome Analysis: A Laboratory Manual 4: Mapping Genomes*, Birren *et al.*, Cold Spring Harbor Press, Cold Spring Harbor, New York (1999); Plant Molecular Biology: A Laboratory Manual, Clark, Springer-Verlag, Berlin, (1997), *Methods in Plant Molecular Biology*, Maliga *et al.*, Cold Spring Harbor Press, Cold Spring Harbor, New York (1995). These texts can, of course, also be referred to in making or using an aspect of the invention. It is understood that any of the agents of the invention can be substantially purified and/or be biologically active and/or recombinant.

Agents

(a) Nucleic Acid Molecules

Agents of the present invention include plant nucleic acid molecules and more preferably include maize, soybean, cotton, sorghum, teosinte, wheat, and rice nucleic acid molecules.

A subset of the nucleic acid molecules of the present invention includes nucleic acid molecules that are marker molecules. Another subset of the nucleic acid molecules of the present invention includes nucleic acid molecules that encode a protein or fragment thereof. Another subset of the nucleic acid molecules of the present invention is cDNA molecules.

Fragment nucleic acid molecules may encode significant portion(s) of, or indeed most of, these nucleic acid molecules. Alternatively, the fragments may comprise smaller oligonucleotides (having from about 15 to about 250 nucleotide residues and more preferably, about 15 to

about 30 nucleotide residues, or more preferably about 30 to about 50 nucleotide residues, or again more preferably about 50 to about 100 nucleotide residues).

The term "substantially purified," as used herein, refers to a molecule separated from substantially all other molecules normally associated with it in its native state. More preferably a substantially purified molecule is the predominant species present in a preparation. A substantially purified molecule may be greater than 60% free, preferably 75% free, more preferably 90% free, and most preferably 95% free from the other molecules (exclusive of solvent) present in the natural mixture. The term "substantially purified" is not intended to encompass molecules present in their native state.

The agents of the present invention will preferably be "biologically active" with respect to either a structural attribute, such as the capacity of a nucleic acid to hybridize to another nucleic acid molecule, or the ability of a protein to be bound by an antibody (or to compete with another molecule for such binding). Alternatively, such an attribute may be catalytic and thus involve the capacity of the agent to mediate a chemical reaction or response.

The agents of the present invention may also be recombinant. As used herein, the term recombinant, refers to a) molecules that are constructed outside of living cells by joining natural or synthetic DNA segments to DNA molecules that can replicate in a living cell or b) molecules that result from the replication or expression of those molecules described above or c) amino acid molecules from different sources which are joined together.

It is understood that the agents of the present invention may be labeled with reagents that facilitate detection of the agent (*e.g.*, fluorescent labels, Prober *et al.*, *Science 238*:336-340 (1987); Albarella *et al.*, EP 144914; chemical labels, Sheldon *et al.*, U.S. Patent 4,582,789; Albarella *et al.*, U.S. Patent 4,563,417; modified bases, Miyoshi *et al.*, EP 119448).

It is further understood, that the present invention provides recombinant bacterial, mammalian, microbial, insect, fungal and plant cells and viral constructs comprising the agents of the present invention

Nucleic acid molecules or fragments thereof of the present invention are capable of specifically hybridizing to other nucleic acid molecules under certain circumstances. As used herein, two nucleic acid molecules are said to be capable of specifically hybridizing to one another if the two molecules are capable of forming an anti-parallel, double-stranded nucleic acid structure. A

nucleic acid molecule is said to be the "complement" of another nucleic acid molecule if they exhibit complete complementarity. As used herein, molecules are said to exhibit "complete complementarity" when every nucleotide of one of the molecules is complementary to a nucleotide of the other. Two molecules are said to be "minimally complementary" if they can hybridize to one another with sufficient stability to permit them to remain annealed to one another under at least conventional "low-stringency" conditions. Similarly, the molecules are said to be "complementary" if they can hybridize to one another with sufficient stability to permit them to remain annealed to one another under conventional "high-stringency" conditions. Conventional stringency conditions are described by Sambrook *et al., Molecular Cloning,* A Laboratory Manual, 2nd Ed., Cold Spring Harbor Press, Cold Spring Harbor, New York (1989) and by Haymes *et al., Nucleic Acid Hybridization, A Practical Approach,* IRL Press, Washington, DC (1985). Departures from complete complementarity are therefore permissible, as long as such departures do not completely preclude the capacity of the molecules to form a double-stranded structure. Thus, in order for a nucleic acid molecule to serve as a primer or probe it need only be sufficiently complementary in sequence to be able to form a stable double-stranded structure under the particular solvent and salt concentrations employed.

Appropriate stringency conditions which promote DNA hybridization, for example, 6.0 X sodium chloride/sodium citrate (SSC) at about 45°C, followed by a wash of 2.0 X SSC at 50°C, are known to those skilled in the art or can be found in *Current Protocols in Molecular Biology,* John Wiley & Sons, N.Y. (1989), 6.3.1-6.3.6. For example, the salt concentration in the wash step can be selected from a low stringency of about 2.0 X SSC at 50°C to a high stringency of about 0.2 X SSC at 50°C. In addition, the temperature in the wash step can be increased from low stringency conditions at room temperature, about 22°C, to high stringency conditions at about 65°C. Both temperature and salt may be varied, or either the temperature or the salt concentration may be held constant while the other variable is changed.

In a preferred embodiment, a nucleic acid of the present invention will specifically hybridize to one or more of the nucleic acid molecules set forth in SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof under moderately stringent conditions, for example at about 2.0 X SSC and about 65°C.

In a particularly preferred embodiment, a nucleic acid of the present invention will include those nucleic acid molecules that specifically hybridize to one or more of the nucleic acid molecules set forth in SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof under high stringency conditions such as 0.2 X SSC and about 65°C.

In one aspect of the present invention, the nucleic acid molecules of the present invention comprise one or more of the nucleic acid sequences set forth in SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof or fragments of either. In another aspect of the present invention, one or more of the nucleic acid molecules of the present invention share between 100% and 90% sequence identity with one or more of the nucleic acid sequences set forth in SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof or fragments of either. In a further aspect of the present invention, one or more of the nucleic acid molecules of the present invention share between 100% and 95% sequence identity with one or more of the nucleic acid sequences set forth in SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof or fragments of either. In a more preferred aspect of the present invention, one or more of the nucleic acid molecules of the present invention share between 100% and 98% sequence identity with one or more of the nucleic acid sequences set forth in SEQ ID NO: 1 through SEQ ID NO: 463,173 complements thereof or fragments of either. In an even more preferred aspect of the present invention, one or more of the nucleic acid molecules of the present invention share between 100% and 99% sequence identity with one or more of the sequences set forth in SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof.

The term "sequence identity" refers to the extent to which two sequences, nucleotide or amino acid, are invariant throughout the portion at which they are aligned. While there exist a number of methods to measure identity between two polynucleotide or polypeptide sequences, the term "sequence identity" is well known to skilled artisans. Methods commonly employed to determine identity between two sequences include, but are not limited to, those disclosed in Guide to Huge Computers, Martin J. Bishop, ed., Academic Press, San Diego, 1994, and Carillo, H., and Lipton, D., SIAM J Applied Math (1988) 48:1073. Methods to determine identity are codified in computer programs. Preferred computer program methods to determine identity between two sequences include, but are not limited to, the BLAST suite of programs publicly available from NCBI and other sources (BLAST Manual, Altschul et al., Natl. Cent. Biotechnol. Inf.,

Natl. Library Med. (NCBI NLM) NIH, Bethesda, Md. 20894; Altschul et al., J. Mol. Biol. 215:403-410 (1990), Pearson et al., Proc. Natl. Acad. Sci. U.S.A. 85:2444-2448 (1988), the FAST programs (Pearson et al., Proc. Natl. Acad. Sci. U.S.A. 85:2444-2448 (1988).the GAP and BESTFIT programs found in the GCG program package,(Madison, WI) and Cross_Match (Phi Green, University of Washington). Another preferred method to determine identity, is by the method of DNASTAR protein alignment protocol using the Jotun-Hein algorithm (Hein et al., Methods Enzymol. 183:626-645 (1990)).

Unless otherwise noted, "percent sequence identity or percent identity" for this invention refers to the value obtained when using the BLAST 2.0 suite of programs with default parameters (Altschul et al., Nucleic Acids Res. 25:3389-3402, 1997; Altschul et al., J. Mol. Bio. 215: 403-410, 1990) Version 2.0 of BLAST allows the introduction of gaps (deletions and insertions) into alignments.

(i) Nucleic Acid Molecules Encoding Proteins or Fragments Thereof

Nucleic acid molecules of the present invention can comprise sequences that encode a protein or fragment thereof. Such proteins or fragments thereof include homologues of known proteins in other organisms.

In a preferred embodiment of the present invention, a maize, soybean, *Arabidopsis,* wheat, cotton, teosinte, or rice protein or fragment thereof of the present invention is a homologue of another plant protein. In another preferred embodiment of the present invention, a maize, soybean, *Arabidopsis,* wheat, cotton, teosinte, or rice protein or fragment thereof of the present invention is a homologue of a fungal protein. In another preferred embodiment of the present invention, a maize, soybean, *Arabidopsis,* wheat, cotton, teosinte, or rice protein of the present invention is a homologue of a mammalian protein. In another preferred embodiment of the present invention, a maize, soybean, *Arabidopsis,* wheat, cotton, teosinte, or rice protein or fragment thereof of the present invention is a homologue of a bacterial protein. In another preferred embodiment of the present invention, a maize, soybean, *Arabidopsis,* wheat, cotton, teosinte, or rice protein or fragment thereof of the present invention is a homologue of a maize protein. In another preferred embodiment of the present invention, a soybean, *Arabidopsis,* wheat, cotton, teosinte, or rice protein or fragment thereof of the present invention is a homologue of a soybean protein. In another preferred embodiment of the present invention, a maize, soybean, *Arabidopsis,* wheat,

cotton, teosinte, or rice protein or fragment thereof of the present invention is a homologue of a cotton protein. In another preferred embodiment of the present invention, a maize, *Arabidopsis,* wheat, cotton, teosinte, or rice protein or fragment thereof of the present invention is a homologue of a wheat protein. In another preferred embodiment of the present invention, a maize, soybean, *Arabidopsis,* cotton, teosinte, or rice protein or fragment thereof of the present invention is a homologue of an *Arabidopsis* protein. In another preferred embodiment of the present invention, a maize, soybean, *Arabidopsis,* wheat, cotton, teosinte, or rice protein or fragment thereof of the present invention is a homologue of a sorghum protein. In another preferred embodiment of the present invention, a maize, soybean, *Arabidopsis,* wheat, cotton, or rice protein or fragment thereof of the present invention is a homologue of a teosinte protein.

In a preferred embodiment of the present invention, the nucleic molecule of the present invention encodes a maize, soybean, *Arabidopsis,* wheat, cotton, teosinte, or rice protein or fragment thereof where a maize, soybean, *Arabidopsis,* wheat, cotton, teosinte, or rice protein exhibits a BLAST E value score of greater than 1E-12, preferably a BLAST E value score of between about 1E-30 and about 1E-12, even more preferably a BLAST probability E value score of greater than 1E-30 with its homologue.

Nucleic acid molecules of the present invention also include non-maize, non-soybean, non-rice, non-wheat, non-*Arabidopsis*, non-sorghum, non-cotton and non-teosinte homologues. Preferred plant sources of homologues are selected from the group consisting of alfalfa, barley, *Brassica*, broccoli, cabbage, citrus, garlic, oat, oilseed rape, onion, canola, flax, an ornamental plant, pea, peanut, pepper, potato, rice, rye, strawberry, sugarcane, sugarbeet, tomato, poplar, pine, fir, eucalyptus, apple, lettuce, lentils, grape, banana, tea, turf grasses, sunflower, oil palm and *Phaseolus*.

In a preferred embodiment, nucleic acid molecules having SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements and fragments of either can be utilized to obtain such homologues.

In another further aspect of the present invention, nucleic acid molecules of the present invention can comprise sequences which differ from those encoding a protein or fragment thereof in SEQ ID NO: 1 through SEQ ID NO: 463,173 due to fact that the different nucleic acid sequence

encodes a protein having one or more conservative amino acid changes. It is understood that codons capable of coding for such conservative amino acid substitutions are known in the art.

It is well known in the art that one or more amino acids in a native sequence can be substituted with another amino acid(s), the charge and polarity of which are similar to that of the native amino acid, *i.e.*, a conservative amino acid substitution, resulting in a silent change. Conserved substitutions for an amino acid within the native polypeptide sequence can be selected from other members of the class to which the naturally occurring amino acid belongs. Amino acids can be divided into the following four groups: (1) acidic amino acids, (2) basic amino acids, (3) neutral polar amino acids, and (4) neutral nonpolar amino acids. Representative amino acids within these various groups include, but are not limited to, (1) acidic (negatively charged) amino acids such as aspartic acid and glutamic acid; (2) basic (positively charged) amino acids such as arginine, histidine, and lysine; (3) neutral polar amino acids such as glycine, serine, threonine, cysteine, cystine, tyrosine, asparagine, and glutamine; and (4) neutral nonpolar (hydrophobic) amino acids such as alanine, leucine, isoleucine, valine, proline, phenylalanine, tryptophan, and methionine.

Conservative amino acid changes within the native polypeptides sequence can be made by substituting one amino acid within one of these groups with another amino acid within the same group. Biologically functional equivalents of the proteins or fragments thereof of the present invention can have ten or fewer conservative amino acid changes, more preferably seven or fewer conservative amino acid changes, and most preferably five or fewer conservative amino acid changes. The encoding nucleotide sequence will thus have corresponding base substitutions, permitting it to encode biologically functional equivalent forms of the proteins or fragments of the present invention.

It is understood that certain amino acids may be substituted for other amino acids in a protein structure without appreciable loss of interactive binding capacity with structures such as, for example, antigen-binding regions of antibodies or binding sites on substrate molecules. Because it is the interactive capacity and nature of a protein that defines that protein's biological functional activity, certain amino acid sequence substitutions can be made in a protein sequence and, of course, its underlying DNA coding sequence and, nevertheless, obtain a protein with like properties. It is thus contemplated by the inventors that various changes may be made in the pep-

tide sequences of the proteins or fragments of the present invention, or corresponding DNA sequences that encode said peptides, without appreciable loss of their biological utility or activity. It is understood that codons capable of coding for such amino acid changes are known in the art.

In making such changes, the hydropathic index of amino acids may be considered. The importance of the hydropathic amino acid index in conferring interactive biological function on a protein is generally understood in the art (Kyte and Doolittle, *J. Mol. Biol. 157*, 105-132 (1982)). It is accepted that the relative hydropathic character of the amino acid contributes to the secondary structure of the resultant protein, which in turn defines the interaction of the protein with other molecules, for example, enzymes, substrates, receptors, DNA, antibodies, antigens, and the like.

Each amino acid has been assigned a hydropathic index on the basis of its hydrophobicity and charge characteristics (Kyte and Doolittle, *J. Mol. Biol. 157*, 105-132 (1982)); these are isoleucine (+4.5), valine (+4.2), leucine (+3.8), phenylalanine (+2.8), cysteine/cystine (+2.5), methionine (+1.9), alanine (+1.8), glycine (-0.4), threonine (-0.7), serine (-0.8), tryptophan (-0.9), tyrosine (-1.3), proline (-1.6), histidine (-3.2), glutamate (-3.5), glutamine (-3.5), aspartate (-3.5), asparagine (-3.5), lysine (-3.9), and arginine (-4.5).

In making such changes, the substitution of amino acids whose hydropathic indices are within ±2 is preferred, those that are within ±1 are particularly preferred, and those within ±0.5 are even more particularly preferred.

It is also understood in the art that the substitution of like amino acids can be made effectively on the basis of hydrophilicity. U.S. Patent 4,554,101 states that the greatest local average hydrophilicity of a protein, as govern by the hydrophilicity of its adjacent amino acids, correlates with a biological property of the protein.

As detailed in U.S. Patent 4,554,101, the following hydrophilicity values have been assigned to amino acid residues: arginine (+3.0), lysine (+3.0), aspartate (+3.0±1), glutamate (+3.0±1), serine (+0.3), asparagine (+0.2), glutamine (+0.2), glycine (0), threonine (-0.4), proline (-0.5±1), alanine (-0.5), histidine (-0.5), cysteine (-1.0), methionine (-1.3), valine (-1.5), leucine (-1.8), isoleucine (-1.8), tyrosine (-2.3), phenylalanine (-2.5), and tryptophan (-3.4).

In making such changes, the substitution of amino acids whose hydrophilicity values are within ±2 is preferred, those that are within ±1 are particularly preferred, and those within ±0.5 are even more particularly preferred.

In a further aspect of the present invention, one or more of the nucleic acid molecules of the present invention differ in nucleic acid sequence from those encoding a protein or fragment thereof set forth in SEQ ID NO: 1 through SEQ ID NO: 463,173 or fragment thereof due to the fact that one or more codons encoding an amino acid has been substituted for a codon that encodes a nonessential substitution of the amino acid originally encoded.

Agents of the invention include nucleic acid molecules that encode at least about a contiguous 10 amino acid region of a protein of the present invention, more preferably at least about a contiguous 25, 40, 50, 100, or 125 amino acid region of a protein of the present invention. In a preferred embodiment the protein is selected from the group consisting of a plant, more preferably a maize, soybean, *Arabidopsis,* wheat, cotton, teosinte, or rice protein.

(ii) Nucleic Acid Molecule Markers and Probes

One aspect of the present invention concerns nucleic acid molecules of the present invention that can act as markers, for example, those nucleic acid molecules SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof or fragments of either that can act as markers or one or more of the marker molecules encoded by other nucleic acid agents of the present invention.

As used herein, a "marker" is an indicator for the presence of at least one polymorphism. A marker is preferably a nucleic acid molecule.

A "nucleic acid marker" as used herein means a nucleic acid molecule that is capable of being a marker for detecting a polymorphism.

In a preferred embodiment, the level, pattern, occurrence and/or absence of a nucleic acid molecule and/or collection of nucleic acid molecules of the present invention is a marker, for example, for a developmental, commercial or non-commercially valuable trait such as yield or an environmental condition or treatment. It is noted that many agronomic traits can affect yield. These include, without limitation, pod position on the plant, number of internodes, incidence of pod shatter, grain size, efficiency of nodulation and nitrogen fixation, efficiency of nutrient assimilation, resistance to biotic and abiotic stress, carbon assimilation, plant architecture, resistance to lodging, percent seed germination, seedling vigor, and juvenile traits.

As used herein, a "collection of nucleic acid molecules" is a population of nucleic acid molecules where at least two of the nucleic acid molecules differ, at least in part, in their nucleic acid sequence. It is understood, that as used herein, an individual species within a collection of

nucleic acid molecules may be physically separate or alternatively not physically separate from one or more other species within the collection of nucleic acid molecules. An example of a situation where individual species may be physically separate but considered a collection of nucleic acid molecules is where more than two species are present on a single support such as a nylon membrane or a glass but occupy a different position on such support. Examples of situations where individual species are physically separate on a support include microarrays.

As used herein, where a collection of nucleic acids is a marker for a particular attribute, the level, pattern, occurrence and/or absence of the nucleic acid molecules associated with the attribute are not required to be the same between species of the collection. For example, the increase in the level of a species when in combination with the decrease in a second species could be diagnostic for a particular attribute.

In an even more preferred embodiment of the present invention, the level, pattern, occurrence and/or absence of a nucleic acid molecule and/or collection of nucleic acid molecules of the present invention is a marker for a biochemical process or activity where the process or activity is preferably selected from photosynthetic activity, carbohydrate metabolism, amino acid synthesis or degradation, plant hormone or other regulatory molecules, phenolic metabolism, and lipid metabolism, and more preferably selected from the group consisting of biosynthesis of tetrapyrroles, phytochrome metabolism, carbon assimilation, glycolysis and gluconeogenesis metabolism, sucrose metabolism, starch metabolism, phosphogluconate metabolism, galactomannan metabolism, raffinose metabolism, complex carbohydrate synthesis/degradation, phytic acid metabolism, methionine biosynthesis, methionine degradation, lysine metabolism, arginine metabolism, proline metabolism, glutamate/glutamine metabolism, aspartate/asparagine metabolism, cytokinin metabolism, gibberellin metabolism, ethylene metabolism, jasmonic acid synthesis metabolism, transcription factors, R-genes, plant proteases, protein kinases, antifungal proteins, nitrogen and sugar transporters, shikimate metabolism, isoflavone metabolism, phenylpropanoid metabolism, isoprenoid metabolism, alpha-oxidation lipid metabolism, and fatty acid metabolism, and even more preferably selected from the group consisting of: glycolysis metabolism, gluconeogenesis metabolism, sucrose metabolism, sucrose catabolism, reductive pentose phosphate cycle, regulation of C3 photosynthesis, C4 pathway carbon assimilation, enzymes involved

in the C4 pathway, carotenoid metabolism, tocopherol metabolism, phytosterol metabolism, brassinoid metabolism, and proline metabolism.

Genetic markers of the invention include "dominant" or "codominant" markers. "Co-dominant markers" reveal the presence of two or more alleles (two per diploid individual) at a locus. "Dominant markers" reveal the presence of only a single allele per locus. The presence of the dominant marker phenotype (e.g., a band of DNA) is an indication that one allele is in either the homozygous or heterozygous condition. The absence of the dominant marker phenotype (*e.g.*, absence of a DNA band) is merely evidence that "some other" undefined allele is present. In the case of populations where individuals are predominantly homozygous and loci are predominately dimorphic, dominant and codominant markers can be equally valuable. As populations become more heterozygous and multi-allelic, codominant markers often become more informative of the genotype than dominant markers. Marker molecules can be, for example, capable of detecting polymorphisms such as single nucleotide polymorphisms (SNPs).

SNPs can be characterized using any of a variety of methods (Botstein *et al.*, *Am. J. Hum. Genet. 32*:314-331 (1980); Konieczny and Ausubel, *Plant J. 4*:403-410 (1993); Myers *et al.*, *Nature 313*:495-498 (1985); Newton *et al.*, *Nucl. Acids Res. 17*:2503-2516 (1989); Wu *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 86*:2757-2760 (1989); Barany, *Proc. Natl. Acad. Sci. (U.S.A.) 88*:189-193 (1991); Labrune *et al.*, *Am. J. Hum. Genet. 48*: 1115-1120 (1991); Kuppuswami *et al.*, *Proc. Natl. Acad. Sci. USA 88*:1143-1147 (1991); Sarkar *et al.*, *Genomics 13*:441-443 (1992); Nikiforov *et al.*, *Nucl. Acids Res. 22*:4167-4175 (1994); Livak *et al.*, *PCR Methods Appl. 4*:357-362 (1995); Livak *et al.*, *Nature Genet. 9*:341-342 (1995); Chen and Kwok, *Nucl. Acids Res. 25*:347-353 (1997); Tyagi *et al.*, *Nature Biotech. 16*: 49-53 (1998); Haff and Smirnov, *Genome Res. 7*: 378-388 (1997); Neff *et al.*, *Plant J. 14*:387-392 (1998)).

Additional markers, such as AFLP markers, RFLP markers and RAPD markers, can be utilized (Walton, *Seed World* 22-29 (July, 1993); Burow and Blake, *Molecular Dissection of Complex Traits*, 13-29, Paterson (ed.), CRC Press, New York (1988)). Another marker type, RAPDs, is developed from DNA amplification with random primers and result from single base changes and insertions/deletions in plant genomes. They are dominant markers with a medium level of polymorphisms and are highly abundant. AFLP markers require using the PCR on a

subset of restriction fragments from extended adapter primers. These markers are both dominant and codominant are highly abundant in genomes and exhibit a medium level of polymorphism.

A PCR probe is a nucleic acid molecule capable of initiating a polymerase activity while in a double-stranded structure to with another nucleic acid. Various methods for determining the structure of PCR probes and PCR techniques exist in the art. Computer generated searches using programs such as Primer3 (www-genome.wi.mit.edu/cgi-bin/primer/primer3.cgi), STSPipeline (www-genome.wi.mit.edu/cgi-bin/www-STS_Pipeline), or GeneUp (Pesole *et al.*, *BioTechniques 25*:112-123 (1998)), for example, can be used to identify potential PCR primers.

It is understood that a fragment of one or more of the nucleic acid molecules of the present invention may be a probe and preferably a PCR probe.

(b) Protein and Peptide Molecules

A class of agents comprises one or more of the protein or peptide molecules encoded by SEQ ID NO: 1 through SEQ ID NO: 463,173 or one or more of the protein or fragment thereof or peptide molecules encoded by other nucleic acid agents of the present invention. As used herein, the term "protein molecule" or "peptide molecule" includes any molecule that comprises five or more amino acids. It is well know in the art that proteins may undergo modification, including post-translational modifications, such as, but not limited to, disulfide bond formation, glycosylation, phosphorylation, or oligomerization. Thus, as used herein, the term "protein molecule" or "peptide molecule" includes any protein molecule that is modified by any biological or non-biological process. The terms "amino acid" and "amino acids" refer to all naturally occurring L-amino acids. This definition is meant to include norleucine, ornithine, homocysteine, and homoserine.

One or more of the protein or fragment of peptide molecules may be produced via chemical synthesis, or more preferably, by expression in a suitable bacterial or eukaryotic host. Suitable methods for expression are described by Sambrook, *et al.*, (In: *Molecular Cloning, A Laboratory Manual, 2nd Edition, Cold Spring Harbor Press*, Cold Spring Harbor, New York (1989)), or similar texts.

A "protein fragment" is a peptide or polypeptide molecule whose amino acid sequence comprises a subset of the amino acid sequence of that protein. A protein or fragment thereof that comprises one or more additional peptide regions not derived from that protein is a "fusion"

protein. Such molecules may be derivatized to contain carbohydrate or other moieties (such as keyhole limpet hemocyanin, etc.). Fusion protein or peptide molecules of the present invention are preferably produced via recombinant means.

Another class of agents comprise protein or peptide molecules encoded by SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof or, fragments or fusions thereof in which non-essential, or not relevant, amino acid residues have been added, replaced, or deleted. An example of such a homologue is the homologue protein of all non- soybean, non-maize, non-rice, non-cotton, non-sorghum, non-teosinte, non-*Arabidopsis* and non-wheat plant species, including but not limited to alfalfa, barley, *Brassica*, broccoli, cabbage, citrus, garlic, oat, oilseed rape, onion, canola, flax, an ornamental plant, pea, peanut, pepper, potato, rye, strawberry, sugarcane, sugarbeet, tomato, poplar, pine, fir, eucalyptus, apple, lettuce, peas, lentils, grape, banana, tea, turf grasses, etc. Particularly preferred non-soybean, non-maize, non-rice, non-cotton, non-sorghum, non-teosinte, non-*Arabidopsis* and non-wheat plants to utilize for the isolation of homologues would include alfalfa, barley, oat, oilseed rape, canola, ornamentals, sugarcane, sugarbeet, tomato, potato, and turf grasses. Such a homologue can be obtained by any of a variety of methods. Most preferably, as indicated above, one or more of the disclosed sequences (SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof) will be used to define a pair of primers that may be used to isolate the homologue-encoding nucleic acid molecules from any desired species. Such molecules can be expressed to yield homologues by recombinant means.

(c) Antibodies

One aspect of the present invention concerns antibodies, single-chain antigen binding molecules, or other proteins that specifically bind to one or more of the protein or peptide molecules of the present invention and their homologues, fusions or fragments. Such antibodies may be used to quantitatively or qualitatively detect the protein or peptide molecules of the present invention. As used herein, an antibody or peptide is said to "specifically bind" to a protein or peptide molecule of the present invention if such binding is not competitively inhibited by the presence of non-related molecules.

Nucleic acid molecules that encode all or part of the protein of the present invention can be expressed, via recombinant means, to yield protein or peptides that can in turn be used to elicit antibodies that are capable of binding the expressed protein or peptide. Such antibodies

may be used in immunoassays for that protein. Such protein-encoding molecules, or their fragments may be a "fusion" molecule (*i.e.*, a part of a larger nucleic acid molecule) such that, upon expression, a fusion protein is produced. It is understood that any of the nucleic acid molecules of the present invention may be expressed, via recombinant means, to yield proteins or peptides encoded by these nucleic acid molecules.

The antibodies that specifically bind proteins and protein fragments of the present invention may be polyclonal or monoclonal and may comprise intact immunoglobulins, or antigen binding portions of immunoglobulins fragments (such as (F(ab'), F(ab')$_2$), or single-chain immunoglobulins producible, for example, via recombinant means. It is understood that practitioners are familiar with the standard resource materials which describe specific conditions and procedures for the construction, manipulation and isolation of antibodies (*see*, for example, Harlow and Lane, In: *Antibodies: A Laboratory Manual,* Cold Spring Harbor Press, Cold Spring Harbor, New York (1988)).

Murine monoclonal antibodies are particularly preferred. BALB/c mice are preferred for this purpose, however, equivalent strains may also be used. The animals are preferably immunized with approximately 25 µg of purified protein (or fragment thereof) that has been emulsified in a suitable adjuvant (such as TiterMax adjuvant (Vaxcel, Norcross, GA)). Immunization is preferably conducted at two intramuscular sites, one intraperitoneal site and one subcutaneous site at the base of the tail. An additional i.v. injection of approximately 25 µg of antigen is preferably given in normal saline three weeks later. After approximately 11 days following the second injection, the mice may be bled and the blood screened for the presence of anti-protein or peptide antibodies. Preferably, a direct binding Enzyme-Linked Immunoassay (ELISA) is employed for this purpose.

More preferably, the mouse having the highest antibody titer is given a third i.v. injection of approximately 25 µg of the same protein or fragment. The splenic leukocytes from this animal may be recovered 3 days later and then permitted to fuse, most preferably, using polyethylene glycol, with cells of a suitable myeloma cell line (such as, for example, the P3X63Ag8.653 myeloma cell line). Hybridoma cells are selected by culturing the cells under "HAT" (hypoxanthine-aminopterin-thymine) selection for about one week. The resulting clones may then be screened for their capacity to produce monoclonal antibodies ("mAbs"), preferably by direct ELISA.

In one embodiment, anti-protein or peptide monoclonal antibodies are isolated using a fusion of a protein or peptide of the present invention, or conjugate of a protein or peptide of the present invention, as immunogens. Thus, for example, a group of mice can be immunized using a fusion protein emulsified in Freund's complete adjuvant (*e.g.* approximately 50 μg of antigen per immunization). At three-week intervals, an identical amount of antigen is emulsified in Freund's incomplete adjuvant and used to immunize the animals. Ten days following the third immunization, serum samples are taken and evaluated for the presence of antibody. If antibody titers are too low, a fourth booster can be employed. Polysera capable of binding the protein or peptide can also be obtained using this method.

In a preferred procedure for obtaining monoclonal antibodies, the spleens of the above-described immunized mice are removed, disrupted and immune splenocytes are isolated over a ficoll gradient. The isolated splenocytes are fused, using polyethylene glycol with BALB/c-derived HGPRT (hypoxanthine guanine phosphoribosyl transferase) deficient P3x63xAg8.653 plasmacytoma cells. The fused cells are plated into 96 well microtiter plates and screened for hybridoma fusion cells by their capacity to grow in culture medium supplemented with hypothanthine, aminopterin and thymidine for approximately 2-3 weeks.

Hybridoma cells that arise from such incubation are preferably screened for their capacity to produce an immunoglobulin that binds to a protein of interest. An indirect ELISA may be used for this purpose. In brief, the supernatants of hybridomas are incubated in microtiter wells that contain immobilized protein. After washing, the titer of bound immunoglobulin can be determined using, for example, a goat anti-mouse antibody conjugated to horseradish peroxidase. After additional washing, the amount of immobilized enzyme is determined (for example through the use of a chromogenic substrate). Such screening is performed as quickly as possible after the identification of the hybridoma in order to ensure that a desired clone is not overgrown by non-secreting neighbor cells. Desirably, the fusion plates are screened several times since the rates of hybridoma growth vary. In a preferred sub-embodiment, a different antigenic form may be used to screen the hybridoma. Thus, for example, the splenocytes may be immunized with one immunogen, but the resulting hybridomas can be screened using a different immunogen. It is understood that any of the protein or peptide molecules of the present invention may be used to raise antibodies.

As discussed below, such antibody molecules or their fragments may be used for diagnostic purposes. Where the antibodies are intended for diagnostic purposes, it may be desirable to derivatize them, for example with a ligand group (such as biotin) or a detectable marker group (such as a fluorescent group, a radioisotope or an enzyme).

The ability to produce antibodies that bind the protein or peptide molecules of the present invention permits the identification of mimetic compounds of those molecules. A "mimetic compound" is a compound that is not that compound, or a fragment of that compound, but which nonetheless exhibits an ability to specifically bind to antibodies directed against that compound.

It is understood that any of the agents of the present invention can be substantially purified and/or be biologically active and/or recombinant.

(d) Plant Constructs and Plant Transformants

One or more of the nucleic acid molecules of the invention may be used in plant transformation or transfection. Exogenous genetic material may be transferred into a plant cell and the plant cell regenerated into a whole, fertile or sterile plant. Exogenous genetic material is any genetic material, whether naturally occurring or otherwise, from any source that is capable of being inserted into any organism. In a preferred embodiment the exogenous genetic material includes a nucleic acid molecule of the present invention, preferably a nucleic acid molecule having a sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof or fragments of either.

Such genetic material may be transferred into either monocotyledons and dicotyledons including, but not limited to maize, soybean, *Arabidopsis,* phaseolus, peanut, alfalfa, wheat, rice, oat, sorghum, rye, tritordeum, millet, fescue, perennial ryegrass, sugarcane, cranberry, papaya, banana, banana, muskmelon, apple, cucumber, dendrobium, gladiolus, chrysanthemum, liliacea, cotton, eucalyptus, sunflower, canola, turfgrass, sugarbeet, coffee and dioscorea (Christou, In: *Particle Bombardment for Genetic Engineering of Plants,* Biotechnology Intelligence Unit. Academic Press, San Diego, California (1996)).

Transfer of a nucleic acid that encodes for a protein can result in overexpression of that protein in a transformed cell or transgenic plant. One or more of the proteins or fragments thereof encoded by nucleic acid molecules of the invention may be overexpressed in a

transformed cell or transformed plant. Such overexpression may be the result of transient or stable transfer of the exogenous genetic material.

Exogenous genetic material may be transferred into a host cell by the use of a DNA vector or construct designed for such a purpose. Design of such a vector is generally within the skill of the art (*See*, *Plant Molecular Biology: A Laboratory Manual*, Clark (ed.), Springier, New York (1997)).

A construct or vector may include a plant promoter to express the protein or protein fragment of choice. A number of promoters, which are active in plant cells, have been described in the literature. These include the nopaline synthase (NOS) promoter (Ebert *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 84*:5745-5749 (1987)), the octopine synthase (OCS) promoter (which are carried on tumor-inducing plasmids of *Agrobacterium tumefaciens*), the caulimovirus promoters such as the cauliflower mosaic virus (CaMV) 19S promoter (Lawton *et al.*, *Plant Mol. Biol. 9*:315-324 (1987)) and the CaMV 35S promoter (Odell *et al.*, *Nature 313*:810-812 (1985)), the figwort mosaic virus 35S-promoter, the light-inducible promoter from the small subunit of ribulose-1,5-bis-phosphate carboxylase (ssRUBISCO), the Adh promoter (Walker *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 84*:6624-6628 (1987)), the sucrose synthase promoter (Yang *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 87*:4144-4148 (1990)), the R gene complex promoter (Chandler *et al.*, *The Plant Cell 1*:1175-1183 (1989)) and the chlorophyll a/b binding protein gene promoter, etc. These promoters have been used to create DNA constructs that have been expressed in plants; *see, e.g.*, PCT publication WO 84/02913. The CaMV 35S promoters are preferred for use in plants. Promoters known or found to cause transcription of DNA in plant cells can be used in the invention.

For the purpose of expression in source tissues of the plant, such as the leaf, seed, root or stem, it is preferred that the promoters utilized have relatively high expression in these specific tissues. Tissue-specific expression of a protein of the present invention is a particularly preferred embodiment. For this purpose, one may choose from a number of promoters for genes with tissue- or cell-specific or -enhanced expression. Examples of such promoters reported in the literature include the chloroplast glutamine synthetase GS2 promoter from pea (Edwards *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 87*:3459-3463 (1990)), the chloroplast fructose-1,6-biphosphatase (FBPase) promoter from wheat (Lloyd *et al.*, *Mol. Gen. Genet. 225*:209-216 (1991)), the nuclear

photosynthetic ST-LS1 promoter from potato (Stockhaus *et al.*, *EMBO J. 8*:2445-2451 (1989)), the serine/threonine kinase (PAL) promoter and the glucoamylase (CHS) promoter from *Arabidopsis thaliana*. Also reported to be active in photosynthetically active tissues are the ribulose-1,5-bisphosphate carboxylase (RbcS) promoter from eastern larch (*Larix laricina*), the promoter for the *cab* gene, cab6, from pine (Yamamoto *et al.*, *Plant Cell Physiol. 35*:773-778 (1994)), the promoter for the Cab-1 gene from wheat (Fejes *et al.*, *Plant Mol. Biol. 15*:921-932 (1990)), the promoter for the CAB-1 gene from spinach (Lubberstedt *et al.*, *Plant Physiol. 104*:997-1006 (1994)), the promoter for the cab1R gene from rice (Luan *et al.*, *Plant Cell. 4*:971-981 (1992)), the pyruvate, orthophosphate dikinase (PPDK) promoter from maize (Matsuoka *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 90*: 9586-9590 (1993)), the promoter for the tobacco Lhcb1*2 gene (Cerdan *et al.*, *Plant Mol. Biol. 33*:245-255 (1997)), the *Arabidopsis thaliana* SUC2 sucrose-H+ symporter promoter (Truernit *et al.*, *Planta. 196*:564-570 (1995)) and the promoter for the thylakoid membrane proteins from spinach (psaD, psaF, psaE, PC, FNR, atpC, atpD, cab, rbcS). Other promoters for the chlorophyll a/b-binding proteins may also be utilized in the invention, such as the promoters for LhcB gene and PsbP gene from white mustard (*Sinapis alba*; Kretsch *et al.*, *Plant Mol. Biol. 28*:219-229 (1995)).

For the purpose of expression in sink tissues of the plant, such as the tuber of the potato plant, the fruit of tomato, or the seed of maize, wheat, rice and barley, it is preferred that the promoters utilized in the invention have relatively high expression in these specific tissues. A number of promoters for genes with tuber-specific or -enhanced expression are known, including the class I patatin promoter (Bevan *et al.*, *EMBO J. 8*:1899-1906 (1986); Jefferson *et al.*, *Plant Mol. Biol. 14*:995-1006 (1990)), the promoter for the potato tuber ADPGPP genes, both the large and small subunits, the sucrose synthase promoter (Salanoubat and Belliard, *Gene 60*:47-56 (1987), Salanoubat and Belliard, *Gene 84*:181-185 (1989)), the promoter for the major tuber proteins including the 22 kd protein complexes and proteinase inhibitors (Hannapel, *Plant Physiol. 101*:703-704 (1993)), the promoter for the granule bound starch synthase gene (GBSS) (Visser *et al.*, *Plant Mol. Biol. 17*:691-699 (1991)) and other class I and II patatins promoters (Koster-Topfer *et al.*, *Mol Gen Genet. 219*:390-396 (1989); Mignery *et al.*, *Gene. 62*:27-44 (1988)).

Other promoters can also be used to express a protein or fragment thereof in specific tissues, such as seeds or fruits. The promoter for β-conglycinin (Chen *et al.*, *Dev. Genet. 10*: 112-

122 (1989)) or other seed-specific promoters such as the napin and phaseolin promoters, can be used. The zeins are a group of storage proteins found in maize endosperm. Genomic clones for zein genes have been isolated (Pedersen *et al.*, *Cell 29*:1015-1026 (1982)) and the promoters from these clones, including the 15 kD, 16 kD, 19 kD, 22 kD, 27 kD and genes, could also be used. Other promoters known to function, for example, in maize include the promoters for the following genes: *waxy, Brittle, Shrunken 2*, Branching enzymes I and II, starch synthases, de-branching enzymes, oleosins, glutelins and sucrose synthases. A particularly preferred promoter for maize endosperm expression is the promoter for the glutelin gene from rice, more particularly the Osgt-1 promoter (Zheng *et al.*, *Mol. Cell Biol. 13*:5829-5842 (1993)). Examples of promoters suitable for expression in wheat include those promoters for the ADPglucose pyrosynthase (ADPGPP) subunits, the granule bound and other starch synthase, the branching and debranching enzymes, the embryogenesis-abundant proteins, the gliadins and the glutenins. Examples of such promoters in rice include those promoters for the ADPGPP subunits, the granule bound and other starch synthase, the branching enzymes, the debranching enzymes, sucrose synthases and the glutelins. A particularly preferred promoter is the promoter for rice glutelin, Osgt-1. Examples of such promoters for barley include those for the ADPGPP subunits, the granule bound and other starch synthase, the branching enzymes, the debranching enzymes, sucrose synthases, the hordeins, the embryo globulins and the aleurone specific proteins.

Root specific promoters may also be used. An example of such a promoter is the promoter for the acid chitinase gene (Samac *et al.*, *Plant Mol. Biol. 25*:587-596 (1994)). Expression in root tissue could also be accomplished by utilizing the root specific subdomains of the CaMV35S promoter that have been identified (Lam *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 86*:7890-7894 (1989)). Other root cell specific promoters include those reported by Conkling *et al.* (Conkling *et al.*, *Plant Physiol. 93*:1203-1211 (1990)).

Additional promoters that may be utilized are described, for example, in U.S. Patent Nos. 5,378,619; 5,391,725; 5,428,147; 5,447,858; 5,608,144; 5,608,144; 5,614,399; 5,633,441; 5,633,435; and 4,633,436. In addition, a tissue specific enhancer may be used (Fromm *et al.*, *The Plant Cell 1*:977-984 (1989)).

Constructs or vectors may also include, with the coding region of interest, a nucleic acid sequence that acts, in whole or in part, to terminate transcription of that region. A number of

such sequences have been isolated, including the Tr7 3' sequence and the NOS 3' sequence (Ingelbrecht *et al., The Plant Cell 1*:671-680 (1989); Bevan *et al., Nucleic Acids Res. 11*:369-385 (1983)).

A vector or construct may also include regulatory elements. Examples of such include the Adh intron 1 (Callis *et al., Genes and Develop. 1*:1183-1200 (1987)), the sucrose synthase intron (Vasil *et al., Plant Physiol. 91*:1575-1579 (1989)) and the TMV omega element (Gallie *et al., The Plant Cell 1*:301-311 (1989)). These and other regulatory elements may be included when appropriate.

A vector or construct may also include a selectable marker. Selectable markers may also be used to select for plants or plant cells that contain the exogenous genetic material. Examples of such include, but are not limited to: a neo gene (Potrykus *et al., Mol. Gen. Genet. 199*:183-188 (1985)), which codes for kanamycin resistance and can be selected for using kanamycin, G418, etc.; a bar gene which codes for bialaphos resistance; a mutant EPSP synthase gene (Hinchee *et al., Bio/Technology 6*:915-922 (1988)) which encodes glyphosate resistance; a nitrilase gene which confers resistance to bromoxynil (Stalker *et al., J. Biol. Chem. 263*:6310-6314 (1988)); a mutant acetolactate synthase gene (ALS) which confers imidazolinone or sulphonylurea resistance (European Patent Application 154,204 (Sept. 11, 1985)); and a methotrexate resistant DHFR gene (Thillet *et al., J. Biol. Chem. 263*:12500-12508 (1988)).

A vector or construct may also include a transit peptide. Incorporation of a suitable chloroplast transit peptide may also be employed (European Patent Application Publication Number 0218571). Translational enhancers may also be incorporated as part of the vector DNA. DNA constructs could contain one or more 5' non-translated leader sequences that may serve to enhance expression of the gene products from the resulting mRNA transcripts. Such sequences may be derived from the promoter selected to express the gene or can be specifically modified to increase translation of the mRNA. Such regions may also be obtained from viral RNAs, from suitable eukaryotic genes, or from a synthetic gene sequence. For a review of optimizing expression of transgenes, see Koziel *et al., Plant Mol. Biol. 32*:393-405 (1996).

A vector or construct may also include a screenable marker. Screenable markers may be used to monitor expression. Exemplary screenable markers include: a β-glucuronidase or uidA gene (GUS) which encodes an enzyme for which various chromogenic substrates are known (Jeff-

erson, *Plant Mol. Biol, Rep. 5:*387-405 (1987); Jefferson *et al., EMBO J. 6:*3901-3907 (1987)); an R-locus gene, which encodes a product that regulates the production of anthocyanin pigments (red color) in plant tissues (Dellaporta *et al.*, Stadler Symposium *11*:263-282 (1988)); a β-lactamase gene (Sutcliffe *et al., Proc. Natl. Acad. Sci. (U.S.A.) 75*:3737-3741 (1978)), a gene which encodes an enzyme for which various chromogenic substrates are known (e.g., PADAC, a chromogenic cephalosporin); a luciferase gene (Ow *et al., Science 234*:856-859 (1986)); a xylE gene (Zukowsky *et al., Proc. Natl. Acad. Sci. (U.S.A.) 80*:1101-1105 (1983)) which encodes a catechol dioxygenase that can convert chromogenic catechols; an α-amylase gene (Ikatu *et al., Bio/Technol. 8*:241-242 (1990)); a tyrosinase gene (Katz *et al., J. Gen. Microbiol. 129*:2703-2714 (1983)) which encodes an enzyme capable of oxidizing tyrosine to DOPA and dopaquinone which in turn condenses to melanin; an α-galactosidase, which will turn a chromogenic α-galactose substrate.

Included within the terms "selectable or screenable marker genes" are also genes that encode a secretable marker whose secretion can be detected as a means of identifying or selecting for transformed cells. Examples include markers that encode a secretable antigen that can be identified by antibody interaction, or even secretable enzymes that can be detected catalytically. Secretable proteins fall into a number of classes, including small, diffusible proteins which are detectable, (*e.g.*, by ELISA), small active enzymes which are detectable in extracellular solution (*e.g.*, α-amylase, β-lactamase, phosphinothricin transferase), or proteins which are inserted or trapped in the cell wall (such as proteins which include a leader sequence such as that found in the expression unit of extension or tobacco PR-S). Other possible selectable and/or screenable marker genes will be apparent to those of skill in the art.

There are many methods for introducing transforming nucleic acid molecules into plant cells. Suitable methods are believed to include virtually any method by which nucleic acid molecules may be introduced into a cell, such as by *Agrobacterium* infection or direct delivery of nucleic acid molecules such as, for example, by PEG-mediated transformation, by electroporation or by acceleration of DNA coated particles, etc (Potrykus, *Ann. Rev. Plant Physiol. Plant Mol. Biol. 42*:205-225 (1991); Vasil, *Plant Mol. Biol. 25*:925-937 (1994)). For example, electroporation has been used to transform maize protoplasts (Fromm *et al., Nature 312*:791-793 (1986)).

Other vector systems suitable for introducing transforming DNA into a host plant cell include but are not limited to binary artificial chromosome (BIBAC) vectors (Hamilton *et al.*,

Gene 200:107-116 (1997)); and transfection with RNA viral vectors (Della-Cioppa *et al., Ann. N.Y. Acad. Sci.* (1996), 792 (Engineering Plants for Commercial Products and Applications), 57-61). Additional vector systems also include plant selectable YAC vectors such as those described in Mullen *et al., Molecular Breeding 4*:449-457 (1988)).

Technology for introduction of DNA into cells is well known to those of skill in the art. Four general methods for delivering a gene into cells have been described: (1) chemical methods (Graham and van der Eb, *Virology* 54:536-539 (1973)); (2) physical methods such as microinjection (Capecchi, *Cell* 22:479-488 (1980)), electroporation (Wong and Neumann, *Biochem. Biophys. Res. Commun. 107*:584-587 (1982); Fromm *et al., Proc. Natl. Acad. Sci. (U.S.A.) 82*:5824-5828 (1985); U.S. Patent No. 5,384,253); and the gene gun (Johnston and Tang, *Methods Cell Biol. 43*:353-365 (1994)); (3) viral vectors (Clapp, *Clin. Perinatol. 20*:155-168 (1993); Lu *et al., J. Exp. Med. 178*:2089-2096 (1993); Eglitis and Anderson, *Biotechniques* 6:608-614 (1988)); and (4) receptor-mediated mechanisms (Curiel *et al., Hum. Gen. Ther. 3*:147-154 (1992), Wagner *et al., Proc. Natl. Acad. Sci. (USA) 89*:6099-6103 (1992)).

Acceleration methods that may be used include, for example, microprojectile bombardment and the like. One example of a method for delivering transforming nucleic acid molecules to plant cells is microprojectile bombardment. This method has been reviewed by Yang and Christou (eds.), *Particle Bombardment Technology for Gene Transfer*, Oxford Press, Oxford, England (1994)). Non-biological particles (microprojectiles) that may be coated with nucleic acids and delivered into cells by a propelling force. Exemplary particles include those comprised of tungsten, gold, platinum and the like.

A particular advantage of microprojectile bombardment, in addition to it being an effective means of reproducibly transforming monocots, is that neither the isolation of protoplasts (Cristou *et al., Plant Physiol. 87*:671-674 (1988)) nor the susceptibility of *Agrobacterium* infection are required. An illustrative embodiment of a method for delivering DNA into maize cells by acceleration is a biolistics α-particle delivery system, which can be used to propel particles coated with DNA through a screen, such as a stainless steel or Nytex screen, onto a filter surface covered with corn cells cultured in suspension. Gordon-Kamm *et al.*, describes the basic procedure for coating tungsten particles with DNA (Gordon-Kamm *et al., Plant Cell* 2:603-618 (1990)). The screen disperses the tungsten nucleic acid particles so that they are not delivered to

the recipient cells in large aggregates. A particle delivery system suitable for use with the invention is the helium acceleration PDS-1000/He gun is available from Bio-Rad Laboratories (Bio-Rad, Hercules, California)(Sanford *et al., Technique 3*:3-16 (1991)).

For the bombardment, cells in suspension may be concentrated on filters. Filters containing the cells to be bombarded are positioned at an appropriate distance below the microprojectile stopping plate. If desired, one or more screens are also positioned between the gun and the cells to be bombarded.

Alternatively, immature embryos or other target cells may be arranged on solid culture medium. The cells to be bombarded are positioned at an appropriate distance below the microprojectile stopping plate. If desired, one or more screens are also positioned between the acceleration device and the cells to be bombarded. Through the use of techniques set forth herein one may obtain up to 1000 or more foci of cells transiently expressing a marker gene. The number of cells in a focus which express the exogenous gene product 48 hours post-bombardment often range from one to ten and average one to three.

In bombardment transformation, one may optimize the pre-bombardment culturing conditions and the bombardment parameters to yield the maximum numbers of stable transformants. Both the physical and biological parameters for bombardment are important in this technology. Physical factors are those that involve manipulating the DNA/microprojectile precipitate or those that affect the flight and velocity of either the macro- or microprojectiles. Biological factors include all steps involved in manipulation of cells before and immediately after bombardment, the osmotic adjustment of target cells to help alleviate the trauma associated with bombardment and also the nature of the transforming DNA, such as linearized DNA or intact supercoiled plasmids. It is believed that pre-bombardment manipulations are especially important for successful transformation of immature embryos.

In another alternative embodiment, plastids can be stably transformed. Methods disclosed for plastid transformation in higher plants include the particle gun delivery of DNA containing a selectable marker and targeting of the DNA to the plastid genome through homologous recombination (Svab *et al., Proc. Natl. Acad. Sci. (U.S.A.) 87*:8526-8530 (1990); Svab and Maliga, *Proc. Natl. Acad. Sci. (U.S.A.) 90*:913-917 (1993); Staub and Maliga, *EMBO J. 12*:601-606 (1993); U.S. Patents 5, 451,513 and 5,545,818).

Accordingly, it is contemplated that one may wish to adjust various aspects of the bombardment parameters in small-scale studies to fully optimize the conditions. One may particularly wish to adjust physical parameters such as gap distance, flight distance, tissue distance and helium pressure. One may also minimize the trauma reduction factors by modifying conditions which influence the physiological state of the recipient cells and which may therefore influence transformation and integration efficiencies. For example, the osmotic state, tissue hydration and the subculture stage or cell cycle of the recipient cells may be adjusted for optimum transformation. The execution of other routine adjustments will be known to those of skill in the art in light of the present disclosure.

Agrobacterium-mediated transfer is a widely applicable system for introducing genes into plant cells because the DNA can be introduced into whole plant tissues, thereby bypassing the need for regeneration of an intact plant from a protoplast. The use of *Agrobacterium*-mediated plant integrating vectors to introduce DNA into plant cells is well known in the art. See, for example the methods described by Fraley *et al., Bio/Technology 3*:629-635 (1985) and Rogers *et al., Methods Enzymol. 153*:253-277 (1987). Further, the integration of the Ti-DNA is a relatively precise process resulting in few rearrangements. The region of DNA to be transferred is defined by the border sequences and intervening DNA is usually inserted into the plant genome as described (Spielmann *et al., Mol. Gen. Genet. 205*:34 (1986)).

Modern *Agrobacterium* transformation vectors are capable of replication in *E. coli* as well as *Agrobacterium*, allowing for convenient manipulations as described (Klee *et al.*, In: *Plant DNA Infectious Agents*, Hohn and Schell (eds.), Springer-Verlag, New York, pp. 179-203 (1985)). Moreover, technological advances in vectors for *Agrobacterium*-mediated gene transfer have improved the arrangement of genes and restriction sites in the vectors to facilitate construction of vectors capable of expressing various polypeptide coding genes. The vectors described have convenient multi-linker regions flanked by a promoter and a polyadenylation site for direct expression of inserted polypeptide coding genes and are suitable for present purposes (Rogers *et al., Methods Enzymol. 153*:253-277 (1987)). In addition, *Agrobacterium* containing both armed and disarmed Ti genes can be used for the transformations. In those plant strains where *Agrobacterium*-mediated transformation is efficient, it is the method of choice because of the facile and defined nature of the gene transfer.

A transgenic plant formed using *Agrobacterium* transformation methods typically contains a single gene on one chromosome. Such transgenic plants can be referred to as being heterozygous for the added gene. More preferred is a transgenic plant that is homozygous for the added structural gene; *i.e.*, a transgenic plant that contains two added genes, one gene at the same locus on each chromosome of a chromosome pair. A homozygous transgenic plant can be obtained by sexually mating (selfing) an independent segregant transgenic plant that contains a single added gene, germinating some of the seed produced and analyzing the resulting plants produced for the gene of interest.

It is also to be understood that two different transgenic plants can also be mated to produce offspring that contain two independently segregating, exogenous genes. Selfing of appropriate progeny can produce plants that are homozygous for both added, exogenous genes that encode a polypeptide of interest. Backcrossing to a parental plant and out-crossing with a non-transgenic plant are also contemplated, as is vegetative propagation.

Transformation of plant protoplasts can be achieved using methods based on calcium phosphate precipitation, polyethylene glycol treatment, electroporation and combinations of these treatments (*See, for example*, Potrykus *et al., Mol. Gen. Genet. 205*:193-200 (1986); Lorz *et al., Mol. Gen. Genet. 199*:178 (1985); Fromm *et al., Nature 319*:791 (1986); Uchimiya *et al., Mol. Gen. Genet. 204*:204 (1986); Marcotte *et al., Nature 335*:454-457 (1988)).

Application of these systems to different plant strains depends upon the ability to regenerate that particular plant strain from protoplasts. Illustrative methods for the regeneration of cereals from protoplasts are described (Fujimura *et al., Plant Tissue Culture Letters* 2:74 (1985); Toriyama *et al., Theor Appl. Genet. 205*:34 (1986); Yamada *et al., Plant Cell Rep. 4*:85 (1986); Abdullah *et al., Biotechnology 4*:1087 (1986)).

To transform plant strains that cannot be successfully regenerated from protoplasts, other ways to introduce DNA into intact cells or tissues can be utilized. For example, regeneration of cereals from immature embryos or explants can be effected as described (Vasil, *Biotechnology 6*:397 (1988)). In addition, "particle gun" or high-velocity microprojectile technology can be utilized (Vasil *et al., Bio/Technology 10*:667 (1992)).

Using the latter technology, DNA is carried through the cell wall and into the cytoplasm on the surface of small metal particles as described (Klein *et al., Nature 328:*70 (1987); Klein *et*

al., Proc. Natl. Acad. Sci. (U.S.A.) 85:8502-8505 (1988); McCabe *et al., Bio/Technology 6:923* (1988)). The metal particles penetrate through several layers of cells and thus allow the transformation of cells within tissue explants.

Other methods of cell transformation can also be used and include but are not limited to introduction of DNA into plants by direct DNA transfer into pollen (Hess *et al., Intern Rev. Cytol. 107*:367 (1987); Luo *et al., Plant Mol Biol. Reporter 6:*165 (1988)), by direct injection of DNA into reproductive organs of a plant (Pena *et al., Nature 325*:274 (1987)), or by direct injection of DNA into the cells of immature embryos followed by the rehydration of desiccated embryos (Neuhaus *et al., Theor. Appl. Genet. 75*:30 (1987)).

The regeneration, development and cultivation of plants from single plant protoplast transformants or from various transformed explants are well known in the art (Weissbach and Weissbach, In: *Methods for Plant Molecular Biology*, Academic Press, San Diego, CA, (1988)). This regeneration and growth process typically includes the steps of selection of transformed cells, culturing those individualized cells through the usual stages of embryonic development through the rooted plantlet stage. Transgenic embryos and seeds are similarly regenerated. The resulting transgenic rooted shoots are thereafter planted in an appropriate plant growth medium such as soil.

The development or regeneration of plants containing the foreign, exogenous gene that encodes a protein of interest is well known in the art. Preferably, the regenerated plants are self-pollinated to provide homozygous transgenic plants. Otherwise, pollen obtained from the regenerated plants is crossed to seed-grown plants of agronomically important lines. Conversely, pollen from plants of these important lines is used to pollinate regenerated plants. A transgenic plant of the invention containing a desired polypeptide is cultivated using methods well known to one skilled in the art.

There are a variety of methods for the regeneration of plants from plant tissue. The particular method of regeneration will depend on the starting plant tissue and the particular plant species to be regenerated.

Methods for transforming dicots, primarily by use of *Agrobacterium tumefaciens* and obtaining transgenic plants have been published for cotton (U.S. Patent No. 5,004,863; U.S. Patent No. 5,159,135; U.S. Patent No. 5,518,908); soybean (U.S. Patent No. 5,569,834; U.S.

Patent No. 5,416,011; McCabe *et. al., Biotechnology 6:*923 (1988); Christou *et al., Plant Physiol. 87*:671-674 (1988)); *Brassica* (U.S. Patent No. 5,463,174); peanut (Cheng *et al., Plant Cell Rep. 15:*653-657 (1996), McKently *et al., Plant Cell Rep. 14:*699-703 (1995)); papaya; and pea (Grant *et al., Plant Cell Rep. 15*:254-258 (1995)).

Transformation of monocotyledons using electroporation, particle bombardment and *Agrobacterium* have also been reported. Transformation and plant regeneration have been achieved in asparagus (Bytebier *et al., Proc. Natl. Acad. Sci. (USA) 84*:5354 (1987)); barley (Wan and Lemaux, *Plant Physiol 104:*37 (1994)); maize (Rhodes *et al., Science 240:*204 (1988); Gordon-Kamm *et al., Plant Cell* 2:603-618 (1990); Fromm *et al., Bio/Technology 8*:833 (1990); Koziel *et al., Bio/Technology 11*:194 (1993); Armstrong *et al., Crop Science 35:*550-557 (1995)); oat (Somers *et al.*; *Bio/Technology 10*:1589 (1992)); orchard grass (Horn *et al., Plant Cell Rep. 7*:469 (1988)); rice (Toriyama *et al., Theor Appl. Genet. 205*:34 (1986); Part *et al., Plant Mol. Biol. 32*:1135-1148 (1996); Abedinia *et al., Aust. J. Plant Physiol. 24:*133-141 (1997); Zhang and Wu, *Theor. Appl. Genet. 76*:835 (1988); Zhang *et al., Plant Cell Rep. 7*:379 (1988); Battraw and Hall, *Plant Sci. 86*:191-202 (1992); Christou *et al., Bio/Technology 9*:957 (1991)); rye (De la Pena *et al., Nature 325*:274 (1987)); sugarcane (Bower and Birch, *Plant J.* 2:409 (1992)); tall fescue (Wang *et al., Bio/Technology 10*:691 (1992)) and wheat (Vasil *et al., Bio/Technology 10*:667 (1992); U.S. Patent No. 5,631,152).

Assays for gene expression based on the transient expression of cloned nucleic acid constructs have been developed by introducing the nucleic acid molecules into plant cells by polyethylene glycol treatment, electroporation, or particle bombardment (Marcotte *et al., Nature 335*:454-457 (1988); Marcotte *et al., Plant Cell 1:*523-532 (1989); McCarty *et al., Cell 66:*895-905 (1991); Hattori *et al., Genes Dev. 6:*609-618 (1992); Goff *et al., EMBO J. 9:*2517-2522 (1990)). Transient expression systems may be used to functionally dissect gene constructs (*see generally*, Mailga *et al., Methods in Plant Molecular Biology,* Cold Spring Harbor Press (1995)).

Any of the nucleic acid molecules of the invention may be introduced into a plant cell in a permanent or transient manner in combination with other genetic elements such as vectors, promoters, enhancers, *etc*. Further, any of the nucleic acid molecules of the invention may be introduced into a plant cell in a manner that allows for overexpression of the protein or fragment thereof encoded by the nucleic acid molecule.

Cosuppression is the reduction in expression levels, usually at the level of RNA, of a particular endogenous gene or gene family by the expression of a homologous sense construct that is capable of transcribing mRNA of the same strandedness as the transcript of the endogenous gene (Napoli *et al., Plant Cell* 2:279-289 (1990); van der Krol *et al., Plant Cell* 2:291-299 (1990)). Cosuppression may result from stable transformation with a single copy nucleic acid molecule that is homologous to a nucleic acid sequence found within the cell (Prolls and Meyer, *Plant J. 2:*465-475 (1992)) or with multiple copies of a nucleic acid molecule that is homologous to a nucleic acid sequence found within the cell (Mittlesten *et al., Mol. Gen. Genet. 244:*325-330 (1994)). Genes, even though different, linked to homologous promoters may result in the cosuppression of the linked genes (Vaucheret, *C.R. Acad. Sci. III 316:*1471-1483 (1993); Flavell, *Proc. Natl. Acad. Sci. (U.S.A.) 91*:3490-3496 (1994)); van Blokland *et al., Plant J.* 6:861-877 (1994); Jorgensen, *Trends Biotechnol.* 8:340-344 (1990); Meins and Kunz, In: *Gene Inactivation and Homologous Recombination in Plants*, Paszkowski (ed.), pp. 335-348, Kluwer Academic, Netherlands (1994)).

It is understood that one or more of the nucleic acids of the invention may be introduced into a plant cell and transcribed using an appropriate promoter with such transcription resulting in the cosuppression of an endogenous protein.

Antisense approaches are a way of preventing or reducing gene function by targeting the genetic material (Mol *et al., FEBS Lett. 268*:427-430 (1990)). The objective of the antisense approach is to use a sequence complementary to the target gene to block its expression and create a mutant cell line or organism in which the level of a single chosen protein is selectively reduced or abolished. Antisense techniques have several advantages over other 'reverse genetic' approaches. The site of inactivation and its developmental effect can be manipulated by the choice of promoter for antisense genes or by the timing of external application or microinjection. Antisense can manipulate its specificity by selecting either unique regions of the target gene or regions where it shares homology to other related genes (Hiatt *et al.,* In: *Genetic Engineering,* Setlow (ed.), Vol. 11, New York: Plenum 49-63 (1989)).

The principle of regulation by antisense RNA is that RNA that is complementary to the target mRNA is introduced into cells, resulting in specific RNA:RNA duplexes being formed by base pairing between the antisense substrate and the target mRNA (Green *et al., Annu. Rev.*

Biochem. 55:569-597 (1986)). Under one embodiment, the process involves the introduction and expression of an antisense gene sequence. Such a sequence is one in which part or all of the normal gene sequences are placed under a promoter in inverted orientation so that the 'wrong' or complementary strand is transcribed into a noncoding antisense RNA that hybridizes with the target mRNA and interferes with its expression (Takayama and Inouye, *Crit. Rev. Biochem. Mol. Biol.* 25:155-184 (1990)). An antisense vector is constructed by standard procedures and introduced into cells by transformation, transfection, electroporation, microinjection, infection, etc. The type of transformation and choice of vector will determine whether expression is transient or stable. The promoter used for the antisense gene may influence the level, timing, tissue, specificity, or inducibility of the antisense inhibition.

It is understood that the activity of a protein in a plant cell may be reduced or depressed by growing a transformed plant cell containing a nucleic acid molecule whose non-transcribed strand encodes a protein or fragment thereof.

Posttranscriptional gene silencing (PTGS) can result in virus immunity or gene silencing in plants. PTGS is induced by dsRNA and is mediated by an RNA-dependent RNA polymerase, present in the cytoplasm, that requires a dsRNA template. The dsRNA is formed by hybridization of complementary transgene mRNAs or complementary regions of the same transcript. Duplex formation can be accomplished by using transcripts from one sense gene and one antisense gene co-located in the plant genome, a single transcript that has self-complementarity, or sense and antisense transcripts from genes brought together by crossing. The dsRNA-dependent RNA polymerase makes a complementary strand from the transgene mRNA and RNAse molecules attach to this complementary strand (cRNA). These cRNA-RNAse molecules hybridize to the endogene mRNA and cleave the single-stranded RNA adjacent to the hybrid. The cleaved single-stranded RNAs are further degraded by other host RNAses because one will lack a capped 5' end and the other will lack a poly(A) tail (Waterhouse *et al., PNAS 95*: 13959-13964 (1998)).

It is understood that one or more of the nucleic acids of the invention may be introduced into a plant cell and transcribed using an appropriate promoter with such transcription resulting in the postranscriptional gene silencing of an endogenous transcript.

Antibodies have been expressed in plants (Hiatt *et al., Nature 342:*76-78 (1989); Conrad and Fielder, *Plant Mol. Biol. 26:*1023-1030 (1994)). Cytoplasmic expression of a scFv (single-chain Fv antibodies) has been reported to delay infection by artichoke mottled crinkle virus. Transgenic plants that express antibodies directed against endogenous proteins may exhibit a physiological effect (Philips *et al., EMBO J. 16:*4489-4496 (1997); Marion-Poll, *Trends in Plant Science 2:*447-448 (1997)). For example, expressed anti-abscissic antibodies have been reported to result in a general perturbation of seed development (Philips *et al., EMBO J. 16:* 4489-4496 (1997)).

Antibodies that are catalytic may also be expressed in plants (abzymes). The principle behind abzymes is that since antibodies may be raised against many molecules, this recognition ability can be directed toward generating antibodies that bind transition states to force a chemical reaction forward (Persidas, *Nature Biotechnology 15:*1313-1315 (1997); Baca *et al., Ann. Rev. Biophys. Biomol. Struct. 26:*461-493 (1997)). The catalytic abilities of abzymes may be enhanced by site directed mutagenesis. Examples of abzymes are, for example, set forth in U.S. Patent No: 5,658,753; U.S. Patent No. 5,632,990; U.S. Patent No. 5,631,137; U.S. Patent 5,602,015; U.S. Patent No. 5,559,538; U.S. Patent No. 5,576,174; U.S. Patent No. 5,500,358; U.S. Patent 5,318,897; U.S. Patent No. 5,298,409; U.S. Patent No. 5,258,289 and U.S. Patent No. 5,194,585.

It is understood that any of the antibodies of the invention may be expressed in plants and that such expression can result in a physiological effect. It is also understood that any of the expressed antibodies may be catalytic.

The present invention also provides for parts of the plants of the present invention. Plant parts, without limitation, include seed, endosperm, ovule and pollen. In a particularly preferred embodiment of the present invention, the plant part is a seed.

Exemplary Uses

Nucleic acid molecules and fragments thereof of the invention may be employed to obtain other nucleic acid molecules from the same species (nucleic acid molecules from maize may be utilized to obtain other nucleic acid molecules from maize). Such nucleic acid molecules include the nucleic acid molecules that encode the complete coding sequence of a protein and promoters and flanking sequences of such molecules. In addition, such nucleic acid molecules

include nucleic acid molecules that encode for other isozymes or gene family members. Such molecules can be readily obtained by using the above-described nucleic acid molecules or fragments thereof to screen cDNA or genomic libraries. Methods for forming such libraries are well known in the art.

Nucleic acid molecules and fragments thereof of the invention may also be employed to obtain nucleic acid homologues. Such homologues include the nucleic acid molecule of other plants or other organisms (*e.g.*, alfalfa, *Arabidopsis*, barley, *Brassica*, broccoli, cabbage, citrus, cotton, garlic, oat, oilseed rape, onion, canola, flax, an ornamental plant, pea, peanut, pepper, potato, rice, rye, sorghum, strawberry, sugarcane, sugarbeet, tomato, wheat, poplar, pine, fir, eucalyptus, apple, lettuce, lentils, grape, banana, tea, turf grasses, sunflower, oil palm, *Phaseolus,* etc.) including the nucleic acid molecules that encode, in whole or in part, protein homologues of other plant species or other organisms, sequences of genetic elements, such as promoters and transcriptional regulatory elements. Such molecules can be readily obtained by using the above-described nucleic acid molecules or fragments thereof to screen cDNA or genomic libraries obtained from such plant species. Methods for forming such libraries are well known in the art. Such homologue molecules may differ in their nucleotide sequences from those found in one or more of SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof because complete complementarity is not needed for stable hybridization. The nucleic acid molecules of the invention therefore also include molecules that, although capable of specifically hybridizing with the nucleic acid molecules may lack "complete complementarity."

Any of a variety of methods may be used to obtain one or more of the above-described nucleic acid molecules (Zamechik *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 83*:4143-4146 (1986); Goodchild *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 85*:5507-5511 (1988); Wickstrom *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 85*:1028-1032 (1988); Holt *et al.*, *Molec. Cell. Biol. 8*:963-973 (1988); Gerwirtz *et al.*, *Science 242*:1303-1306 (1988); Anfossi *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 86*:3379-3383 (1989); Becker *et al.*, *EMBO J. 8*:3685-3691 (1989)). Automated nucleic acid synthesizers may be employed for this purpose. In lieu of such synthesis, the disclosed nucleic acid molecules may be used to define a pair of primers that can be used with the polymerase chain reaction (Mullis *et al.*, *Cold Spring Harbor Symp. Quant. Biol. 51*:263-273 (1986); Erlich *et al.*, European Patent 50,424; European Patent 84,796; European Patent 258,017; European

Patent 237,362; Mullis, European Patent 201,184; Mullis *et al.*, U.S. Patent 4,683,202; Erlich, U.S. Patent 4,582,788; and Saiki *et al.*, U.S. Patent 4,683,194) to amplify and obtain any desired nucleic acid molecule or fragment.

Promoter sequences and other genetic elements, including but not limited to transcriptional regulatory flanking sequences, associated with one or more of the disclosed nucleic acid sequences can also be obtained using the disclosed nucleic acid sequence provided herein. In one embodiment, such sequences are obtained by incubating nucleic acid molecules of the present invention with members of genomic libraries and recovering clones that hybridize to such nucleic acid molecules thereof. In a second embodiment, methods of "chromosome walking," or inverse PCR may be used to obtain such sequences (Frohman *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 85*:8998-9002 (1988); Ohara *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 86*:5673-5677 (1989); Pang *et al.*, *Biotechniques 22:*1046-1048 (1977); Huang *et al.*, *Methods Mol. Biol. 69:*89-96 (1997); Huang *et al.*, *Method Mol. Biol. 67:*287-294 (1997); Benkel *et al.*, *Genet. Anal. 13:*123-127 (1996); Hartl *et al.*, *Methods Mol. Biol. 58:*293-301 (1996)). The term "chromosome walking" means a process of extending a genetic map by successive hybridization steps.

The nucleic acid molecules of the invention may be used to isolate promoters of cell enhanced, cell specific, tissue enhanced, tissue specific, developmentally or environmentally regulated expression profiles. Isolation and functional analysis of the 5' flanking promoter sequences of these genes from genomic libraries, for example, using genomic screening methods and PCR techniques would result in the isolation of useful promoters and transcriptional regulatory elements. These methods are known to those of skill in the art and have been described (See, for example, Birren *et al.*, *Genome Analysis*: *Analyzing DNA*, 1, (1997), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.). Promoters obtained utilizing the nucleic acid molecules of the invention could also be modified to affect their control characteristics. Examples of such modifications would include but are not limited to enhancer sequences. Such genetic elements could be used to enhance gene expression of new and existing traits for crop improvement.

Another subset of the nucleic acid molecules of the invention includes nucleic acid molecules that are markers. The markers can be used in a number of conventional ways in the

field of molecular genetics. Such markers include nucleic acid molecules SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof or fragments of either that can act as markers and other nucleic acid molecules of the present invention that can act as markers.

The genomes of animals and plants naturally undergo spontaneous mutation in the course of their continuing evolution (Gusella, *Ann. Rev. Biochem. 55*:831-854 (1986)). A "polymorphism" is a variation or difference in the sequence of the gene or its flanking regions that arises in some of the members of a species. The variant sequence and the "original" sequence co-exist in the species' population. In some instances, such co-existence is in stable or quasi-stable equilibrium.

A polymorphism is thus said to be "allelic," in that, due to the existence of the polymorphism, some members of a species may have the original sequence (i.e., the original "allele") whereas other members may have the variant sequence (i.e., the variant "allele"). In the simplest case, only one variant sequence may exist and the polymorphism is thus said to be di-allelic. In other cases, the species' population may contain multiple alleles and the polymorphism is termed tri-allelic, etc. A single gene may have multiple different unrelated polymorphisms. For example, it may have a di-allelic polymorphism at one site and a multi-allelic polymorphism at another site.

The variation that defines the polymorphism may range from a single nucleotide variation to the insertion or deletion of extended regions within a gene. In some cases, the DNA sequence variations are in regions of the genome that are characterized by short tandem repeats (STRs) that include tandem di- or tri-nucleotide repeated motifs of nucleotides. Polymorphisms characterized by such tandem repeats are referred to as "variable number tandem repeat" ("VNTR") polymorphisms. VNTRs have been used in identity analysis (Weber, U.S. Patent 5,075,217; Armour *et al., FEBS Lett. 307*:113-115 (1992); Jones *et al., Eur. J. Haematol. 39*:144-147 (1987); Horn *et al.*, PCT Patent Application WO91/14003; Jeffreys, European Patent Application 370,719; Jeffreys, U.S. Patent 5,175,082; Jeffreys *et al., Amer. J. Hum. Genet. 39*:11-24 (1986); Jeffreys *et al., Nature 316*:76-79 (1985); Gray *et al., Proc. R. Acad. Soc. Lond. 243*:241-253 (1991); Moore *et al., Genomics 10*:654-660 (1991); Jeffreys *et al., Anim. Genet. 18*:1-15 (1987); Hillel *et al., Anim. Genet. 20*:145-155 (1989); Hillel *et al., Genet. 124*:783-789 (1990)).

The detection of polymorphic sites in a sample of DNA may be facilitated through the use of nucleic acid amplification methods. Such methods specifically increase the concentration of polynucleotides that span the polymorphic site, or include that site and sequences located either distal or proximal to it. Such amplified molecules can be readily detected by gel electrophoresis or other means.

In an alternative embodiment, such polymorphisms can be detected through the use of a marker nucleic acid molecule that is physically linked to such polymorphism(s). For this purpose, marker nucleic acid molecules comprising a nucleotide sequence of a polynucleotide located within 1 mb of the polymorphism(s) and more preferably within 100kb of the polymorphism(s) and most preferably within 10kb of the polymorphism(s) can be employed.

The identification of a polymorphism can be determined in a variety of ways. By correlating the presence or absence of it in a plant with the presence or absence of a phenotype, it is possible to predict the phenotype of that plant. If a polymorphism creates or destroys a restriction endonuclease cleavage site, or if it results in the loss or insertion of DNA (e.g., a VNTR polymorphism), it will alter the size or profile of the DNA fragments that are generated by digestion with that restriction endonuclease. As such, individuals that possess a variant sequence can be distinguished from those having the original sequence by restriction fragment analysis. Polymorphisms that can be identified in this manner are termed "restriction fragment length polymorphisms" ("RFLPs") (Glassberg, UK Patent Application 2135774; Skolnick *et al.*, *Cytogen. Cell Genet. 32*:58-67 (1982); Botstein *et al.*, *Ann. J. Hum. Genet. 32*:314-331 (1980); Fischer *et al.*, (PCT Application WO90/13668; Uhlen, PCT Application WO90/11369).

Polymorphisms can also be identified by Single Strand Conformation Polymorphism (SSCP) analysis (Elles, *Methods in Molecular Medicine: Molecular Diagnosis of Genetic Diseases*, Humana Press (1996)); Orita *et al., Genomics 5:*874-879 (1989)). A number of protocols have been described for SSCP including, but not limited to, Lee *et al., Anal. Biochem. 205:*289-293 (1992); Suzuki *et al., Anal. Biochem. 192:*82-84 (1991); Lo *et al., Nucleic Acids Research 20:*1005-1009 (1992); Sarkar *et al., Genomics 13:*441-443 (1992). It is understood that one or more of the nucleic acids of the invention, may be utilized as markers or probes to detect polymorphisms by SSCP analysis.

Polymorphisms may also be found using a DNA fingerprinting technique called amplified fragment length polymorphism (AFLP), which is based on the selective PCR amplification of restriction fragments from a total digest of genomic DNA to profile that DNA (Vos *et al.*, *Nucleic Acids Res. 23*:4407-4414 (1995)). This method allows for the specific co-amplification of high numbers of restriction fragments, which can be visualized by PCR without knowledge of the nucleic acid sequence. It is understood that one or more of the nucleic acids of the invention, may be utilized as markers or probes to detect polymorphisms by AFLP analysis or for fingerprinting RNA.

Polymorphisms may also be found using random amplified polymorphic DNA (RAPD) (Williams *et al., Nucl. Acids Res. 18*:6531-6535 (1990)) and cleaveable amplified polymorphic sequences (CAPS) (Lyamichev *et al.*, *Science 260*:778-783 (1993)). It is understood that one or more of the nucleic acid molecules of the invention, may be utilized as markers or probes to detect polymorphisms by RAPD or CAPS analysis.

Through genetic mapping, a fine scale linkage map can be developed using DNA markers and, then, a genomic DNA library of large-sized fragments can be screened with molecular markers linked to the desired trait. Molecular markers are advantageous for agronomic traits that are otherwise difficult to tag, such as resistance to pathogens, insects and nematodes, tolerance to abiotic stress, quality parameters and quantitative traits such as high yield potential.

Requirements for marker-assisted selection in a plant breeding program often are: (1) the marker(s) should co-segregate or be closely linked with the desired trait; (2) an efficient means of screening large populations for the molecular marker(s) should be available; and (3) the screening technique should have high reproducibility across laboratories and preferably be economical to use and be user-friendly.

The genetic linkage of marker molecules can be established by a gene mapping model such as, without limitation, the flanking marker model reported by Lander and Botstein, *Genetics 121*:185-199 (1989) and the interval mapping, based on maximum likelihood methods described by Lander and Botstein, *Genetics 121*:185-199 (1989) and implemented in the software package MAPMAKER/QTL (Lincoln and Lander, *Mapping Genes Controlling Quantitative Traits Using MAPMAKER/QTL,* Whitehead Institute for Biomedical Research, Massachusetts, (1990). Additional software includes Qgene, Version 2.23 (1996), Department of Plant Breeding and

Biometry, 266 Emerson Hall, Cornell University, Ithaca, NY). Use of Qgene software is a particularly preferred approach.

A maximum likelihood estimate (MLE) for the presence of a marker is calculated, together with an MLE assuming no QTL effect, to avoid false positives. A $\log_{10}$ of an odds ratio (LOD) is then calculated as: LOD = $\log_{10}$ (MLE for the presence of a QTL/MLE given no linked QTL).

The LOD score essentially indicates how much more likely the data are to have arisen assuming the presence of a QTL than in its absence. The LOD threshold value for avoiding a false positive with a given confidence, say 95%, depends on the number of markers and the length of the genome. Graphs indicating LOD thresholds are set forth in Lander and Botstein, *Genetics 121*:185-199 (1989) and further described by Arús and Moreno-González , *Plant Breeding*, Hayward *et al.,* (eds.) Chapman & Hall, London, pp. 314-331 (1993).

Additional models can be used. Many modifications and alternative approaches to interval mapping have been reported, including the use non-parametric methods (Kruglyak and Lander, *Genetics 139:*1421-1428 (1995)). Multiple regression methods or models can be also be used, in which the trait is regressed on a large number of markers (Jansen, *Biometrics in Plant Breeding*, van Oijen and Jansen (eds.), Proceedings of the Ninth Meeting of the Eucarpia Section Biometrics in Plant Breeding, The Netherlands, pp. 116-124 (1994); Weber and Wricke, *Advances in Plant Breeding,* Blackwell, Berlin, 16 (1994)). Procedures combining interval mapping with regression analysis, whereby the phenotype is regressed onto a single putative QTL at a given marker interval and at the same time onto a number of markers that serve as 'cofactors,' have been reported by Jansen and Stam, *Genetics 136*:1447-1455 (1994), and Zeng, *Genetics 136*:1457-1468 (1994). Generally, the use of cofactors reduces the bias and sampling error of the estimated QTL positions (Utz and Melchinger, *Biometrics in Plant Breeding,* van Oijen and Jansen (eds.) Proceedings of the Ninth Meeting of the Eucarpia Section Biometrics in Plant Breeding, The Netherlands, pp.195-204 (1994), thereby improving the precision and efficiency of QTL mapping (Zeng, *Genetics 136*:1457-1468 (1994)). These models can be extended to multi-environment experiments to analyze genotype-environment interactions (Jansen *et al., Theo. Appl. Genet. 91:*33-37 (1995)).

It is understood that one or more of the nucleic acid molecules of the invention may be used as molecular markers. It is also understood that one or more of the protein molecules of the invention may be used as molecular markers.

In accordance with this aspect of the invention, a sample nucleic acid is obtained from plant cells or tissues. Any source of nucleic acid may be used. Preferably, the nucleic acid is genomic DNA. The nucleic acid is subjected to restriction endonuclease digestion. For example, one or more nucleic acid molecule or fragment thereof of the invention can be used as a probe in accordance with the above-described polymorphic methods. The polymorphism obtained in this approach can then be cloned to identify the mutation at the coding region, which alters structure, or regulatory region of the gene, which affects its expression level.

In an aspect of the present invention, one or more of the nucleic molecules of the present invention are used to determine the level (*i.e.*, the concentration of mRNA in a sample, *etc.*) in a plant (preferably maize or soybean) or pattern (*i.e.*, the kinetics of expression, rate of decomposition, stability profile, *etc.*) of the expression of a protein encoded in part or whole by one or more of the nucleic acid molecule of the present invention (collectively, the "Expression Response" of a cell or tissue).

As used herein, the Expression Response manifested by a cell or tissue is said to be "altered" if it differs from the Expression Response of cells or tissues of plants not exhibiting the phenotype. To determine whether a Expression Response is altered, the Expression Response manifested by the cell or tissue of the plant exhibiting the phenotype is compared with that of a similar cell or tissue sample of a plant not exhibiting the phenotype. As will be appreciated, it is not necessary to re-determine the Expression Response of the cell or tissue sample of plants not exhibiting the phenotype each time such a comparison is made; rather, the Expression Response of a particular plant may be compared with previously obtained values of normal plants. As used herein, the phenotype of the organism is any of one or more characteristics of an organism (*e.g.* disease resistance, pest tolerance, environmental tolerance such as tolerance to abiotic stress, male sterility, quality improvement or yield *etc.*). A change in genotype or phenotype may be transient or permanent. Also as used herein, a tissue sample is any sample that comprises more than one cell. In a preferred aspect, a tissue sample comprises cells that share a common characteristic (e.g. derived from root, seed, flower, leaf, stem or pollen etc.).

In one aspect of the present invention, an evaluation can be conducted to determine whether a particular mRNA molecule is present. One or more of the nucleic acid molecules of the present invention are utilized to detect the presence or quantity of the mRNA species. Such molecules are then incubated with cell or tissue extracts of a plant under conditions sufficient to permit nucleic acid hybridization. The detection of double-stranded probe-mRNA hybrid molecules is indicative of the presence of the mRNA; the amount of such hybrid formed is proportional to the amount of mRNA. Thus, such probes may be used to ascertain the level and extent of the mRNA production in a plant's cells or tissues. Such nucleic acid hybridization may be conducted under quantitative conditions (thereby providing a numerical value of the amount of the mRNA present). Alternatively, the assay may be conducted as a qualitative assay that indicates either that the mRNA is present, or that its level exceeds a user set, predefined value.

A number of methods can be used to compare the expression response between two or more samples of cells or tissue. These methods include hybridization assays, such as Northerns, RNAse protection assays, and *in situ* hybridization. Alternatively, the methods include PCR-type assays. In a preferred method, the expression response is compared by hybridizing nucleic acids from the two or more samples to an array of nucleic acids. The array contains a plurality of suspected sequences known or suspected of being present in the cells or tissue of the samples.

An advantage of *in situ* hybridization over more conventional techniques for the detection of nucleic acids is that it allows an investigator to determine the precise spatial population (Angerer *et al., Dev. Biol. 101*:477-484 (1984); Angerer *et al., Dev. Biol. 112*:157-166 (1985); Dixon *et al., EMBO J. 10*:1317-1324 (1991)). *In situ* hybridization may be used to measure the steady-state level of RNA accumulation (Hardin *et al., J. Mol. Biol. 202*:417-431 (1989)). A number of protocols have been devised for *in situ* hybridization, each with tissue preparation, hybridization and washing conditions (Meyerowitz, *Plant Mol. Biol. Rep. 5*:242-250 (1987); Cox and Goldberg, In: *Plant Molecular Biology: A Practical Approach,* Shaw (ed.), pp. 1-35, IRL Press, Oxford (1988); Raikhel *et al., In situ RNA hybridization in plant tissues*, In: *Plant Molecular Biology Manual*, vol. B9:1-32, Kluwer Academic Publisher, Dordrecht, Belgium (1989)).

In situ hybridization also allows for the localization of proteins within a tissue or cell (Wilkinson, *In Situ Hybridization,* Oxford University Press, Oxford (1992); Langdale, *In Situ*

Hybridization In: *The Maize Handbook,* Freeling and Walbot (eds.), pp. 165-179, Springer-Verlag, New York (1994)). It is understood that one or more of the molecules of the invention, preferably one or more of the nucleic acid molecules or fragments thereof of the invention or one or more of the antibodies of the invention may be utilized to detect the level or pattern of a protein or mRNA thereof by *in situ* hybridization.

Fluorescent *in situ* hybridization allows the localization of a particular DNA sequence along a chromosome which is useful, among other uses, for gene mapping, following chromosomes in hybrid lines or detecting chromosomes with translocations, transversions or deletions. *In situ* hybridization has been used to identify chromosomes in several plant species (Griffor *et al., Plant Mol. Biol. 17*:101-109 (1991); Gustafson *et al., Proc. Natl. Acad. Sci. (U.S.A.) 87*:1899-1902 (1990); Mukai and Gill, *Genome 34*:448-452 (1991); Schwarzacher and Heslop-Harrison, *Genome 34*:317-323 (1991); Wang *et al., Jpn. J. Genet. 66*:313-316 (1991); Parra and Windle, *Nature Genetics 5*:17-21 (1993)). It is understood that the nucleic acid molecules of the invention may be used as probes or markers to localize sequences along a chromosome.

Another method to localize the expression of a molecule is tissue printing. Tissue printing provides a way to screen, at the same time on the same membrane many tissue sections from different plants or different developmental stages (Yomo and Taylor, *Planta 112*:35-43 (1973); Harris and Chrispeels, *Plant Physiol.* 56:292-299 (1975); Cassab and Varner, *J. Cell. Biol. 105*:2581-2588 (1987); Spruce *et al., Phytochemistry 26:*2901-2903 (1987); Barres *et al., Neuron 5:*527-544 (1990); Reid and Pont-Lezica, *Tissue Printing: Tools for the Study of Anatomy, Histochemistry and Gene Expression,* Academic Press, New York, New York (1992); Reid *et al., Plant Physiol. 93:*160-165 (1990); Ye *et al., Plant J. 1:*175-183 (1991)).

It is understood that one or more of the molecules of the invention, preferably one or more of the nucleic acid molecules of the present invention or one or more of the antibodies of the invention may be utilized to detect the presence or quantity of a protein or fragment of the invention by tissue printing.

Further it is also understood that any of the nucleic acid molecules of the invention may be used as marker nucleic acids and or probes in connection with methods that require probes or marker nucleic acids. As used herein, a probe is an agent that is utilized to determine an attribute

or feature (e.g. presence or absence, location, correlation, etc.) of a molecule, cell, tissue or plant. As used herein, a marker nucleic acid is a nucleic acid molecule that is utilized to determine an attribute or feature (*e.g.,* presence or absence, location, correlation, etc.) or a molecule, cell, tissue or plant.

A microarray-based method for high-throughput monitoring of gene expression may be utilized to measure expression response Schena *et al., Science 270*:467-470 (1995); http://cmgm.stanford.edu/pbrown/array.html; Shalon, Ph.D. Thesis, Stanford University (1996). This approach is based on using arrays of DNA targets (e.g. cDNA inserts, colonies, or polymerase chain reaction products) for hybridization to a "complex probe" prepared with RNA extracted from a given cell line or tissue. The probe may be produced by reverse transcription of mRNA or total RNA and labeled with radioactive or fluorescent labeling. The probe is complex in that it contains many different sequences in various amounts, corresponding to the numbers of copies of the original mRNA species extracted from the sample.

The initial RNA source will typically be derived from a physiological source. The physiological source may be derived from a variety of eukaryotic sources, with physiological sources of interest including sources derived from single celled organisms such as yeast and multicellular organisms, including plants and animals, particularly plants, where the physiological sources from multicellular organisms may be derived from particular organs or tissues of the multicellular organism, or from isolated cells derived therefrom. The physiological sources may be derived from multicellular organisms at different developmental stages (*e.g.*, 10-day-old seedlings), grown under different environmental conditions (*e.g.*, drought-stressed plants) or treated with chemicals.

In obtaining the sample of RNAs to be analyzed from the physiological source from which it is derived, the physiological source may be subjected to a number of different processing steps, where such processing steps might include tissue homogenation, cell isolation and cytoplasmic extraction, nucleic acid extraction and the like, where such processing steps are known to the those of skill in the art. Methods of isolating RNA from cells, tissues, organs or whole organisms are known to those of skill in the art and are described in Maniatis et al., Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Press) (1989).

The DNA may be placed on nylon or glass "microarrays" regularly arranged with a spot spacing of 1 mm or less. Expression levels can be measured for hundreds or thousands of genes, by using less than 2 micrograms of polyA+ RNA and determining the relative mRNA abundances down to one in ten thousand or less (Granjeaud *et. al., BioEssays* 21:781-790 (1999)).

In addition to arrays of cDNA clones or inserts, arrays of oligonucleotides are also used to study differential gene expression. In an oligonucleotide array, the genes of interest are represented by a series of approximately 20 nucleotide oligomers that are unique to each gene. Labeled mRNA is prepared and hybridization signals are detected from specific sets of oligos that represent different genes supplemented by a set of control oligonucleotides. Potential advantages of the oligonucleotide array include enhanced specificity and sensitivity through the parallel analysis of "perfect match" oligos and "mismatch" oligos for each gene. The hybridization conditions can be adjusted to distinguish a perfect heteroduplex from a single base mismatch, thus allowing subtraction of nonspecific hybridization signals from specific hybridization signals. A disadvantage of oligonucleotide arrays relative to cDNA arrays is the limitation of the technology to genes of known sequence (Granjeaud *et. al., BioEssays* 21:781-790 (1991); Carulli *et al., Journal of Cellular Biochemistry Supplements 30/31*:286-296 (1998)).

These techniques have been successfully used to characterize patterns of gene expression associated with, for example, various important physiological changes in yeast, including the mitotic cell cycle, the heat shock response, and comparison between mating types. Once a set of comparable expression profiles is obtained, e.g. for cells at different time points or at different cellular states, a clustering algorithm generally is used to group sets of genes which share similar expression patterns. The clusters obtained can then be analyzed in the light of available functional annotations, often leading to associations of poorly characterized genes with genes whose function and regulation are better understood.

Regulatory networks that control gene expression can be characterized using microarray technology (DeRisi *et al., Science 278:* 680-686 (1997); Winzler *et al. Science 28:* 1194-1197 (1998); Cho *et al. Mol Cell 2:* 65-73 (1998); Spellman *et al. Mol Biol Cell 95:* 14863-14868 (1998). For example, it is has been reported that both cDNA and oligonucleotide arrays have been used to monitor gene expression in synchronized cell cultures. Analysis of the

corresponding temporal patterns of gene expression resulted in the identification of over 400 cell cycle-regulated genes. In order to identify possible common regulatory mechanisms accounting for co-expression, consensus motifs in putative regulatory sequences upstream of the corresponding ORFs were examined. This resulted in the identification of several new potential binding sites for known factors or complexes involved in the coordinated transcription of genes during specific phases of the cell cycle (Thieffry, D. *BioEssays 21*: 895-899 (1999)).

The microarray approach may be used with polypeptide targets (U.S. Patent No. 5,445,934; U.S. Patent No: 5,143,854; U.S. Patent No. 5,079,600; U.S. Patent No. 4,923,901) synthesized on a substrate (microarray) and these polypeptides can be screened with either (Fodor *et al.*, *Science 251*:767-773 (1991)). It is understood that one or more of the nucleic acid molecules or protein or fragments thereof of the invention may be utilized in a microarray-based method.

In a preferred embodiment of the present invention microarrays may be prepared that comprise nucleic acid molecules where preferably at least 10%, preferably at least 25%, more preferably at least 50% and even more preferably at least 75%, 80%, 85%, 90% or 95% of the nucleic acid molecules located on that array are selected from the group of nucleic acid molecules that specifically hybridize to one or more nucleic acid molecule having a nucleic acid sequence selected from the group of SEQ ID NO: 1 through SEQ ID NO: 463,173 or complement thereof or fragments of either.

In another preferred embodiment of the present invention microarrays may be prepared that comprise nucleic acid molecules where preferably at least 10%, preferably at least 25%, more preferably at least 50% and even more preferably at least 75%, 80%, 85%, 90% or 95% of the nucleic acid molecules located on that array are selected from the group of nucleic acid molecules having a nucleic acid sequence selected from the group of SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof.

In an even more preferred embodiment of the present invention, the microarray comprises a nucleic acid molecule and/or collection of nucleic acid molecules of the present invention where the nucleic acid molecule and/or collection of nucleic acid molecules are capable of determining or predicting a component or attribute of a biochemical process or activity where the process or activity is preferably selected from photosynthetic activity,

carbohydrate metabolism, amino acid synthesis or degradation, plant hormone or other regulatory molecules, phenolic metabolism, and lipid metabolism, and more preferably selected from the group consisting of biosynthesis of tetrapyrroles, phytochrome metabolism, carbon assimilation, glycolysis and gluconeogenesis metabolism, sucrose metabolism, starch metabolism, phosphogluconate metabolism, galactomannan metabolism, raffinose metabolism, complex carbohydrate synthesis/degradation, phytic acid metabolism, methionine biosynthesis, methionine degradation, lysine metabolism, arginine metabolism, proline metabolism, glutamate/glutamine metabolism, aspartate/asparagine metabolism, cytokinin metabolism, gibberellin metabolism, ethylene metabolism, jasmonic acid synthesis metabolism, transcription factors, R-genes, plant proteases, protein kinases, antifungal proteins, nitrogen transporters, sugar transporters, shikimate metabolism, isoflavone metabolism, phenylpropanoid metabolism, isoprenoid metabolism, alpha-oxidation lipid metabolism, and fatty acid metabolism, and even more preferably selected from the group consisting of: glycolysis metabolism, gluconeogenesis metabolism, sucrose metabolism, sucrose catabolism, reductive pentose phosphate cycle, regulation of C3 photosynthesis, C4 pathway carbon assimilation, enzymes involved in the C4 pathway, carotenoid metabolism, tocopherol metabolism, phytosterol metabolism, brassinoid metabolism, and proline metabolism.

In an even more preferred embodiment of the present invention, the microarray comprises a nucleic acid molecule and/or collection of nucleic acid molecules of the present invention where the nucleic acid molecule and/or collection of nucleic acid molecules are capable of detecting or predicting a component or attribute of at least two, more preferable at least three, four, five, six, seven, eight, nine, ten, eleven, twelve, thirteen, fourteen, fifteen, sixteen, seventeen, eighteen, nineteen, twenty, twenty one, twenty two, twenty three, twenty four, twenty five, twenty six, twenty seven, twenty eight, twenty nine, thirty, thirty one, thirty two, thirty three, thirty four, thirty five, thirty six, thirty seven, thirty eight, thirty nine, forty, forty one, forty two, forty three, forty four, forty five or forty six biochemical processes or activities where the biochemical processes or activities are selected from the following: photosynthetic activity, carbohydrate metabolism, amino acid synthesis or degradation, plant hormone or other regulatory molecules, phenolic metabolism, lipid metabolism, biosynthesis of tetrapyrroles, phytochrome metabolism, carbon assimilation, glycolysis and gluconeogenesis

metabolism, sucrose metabolism, starch metabolism, phosphogluconate metabolism, galactomannan metabolism, raffinose metabolism, complex carbohydrate synthesis/degradation, phytic acid metabolism, methionine biosynthesis, methionine degradation, lysine metabolism, arginine metabolism, proline metabolism, glutamate/glutamine, aspartate/asparagine metabolism, cytokinin metabolism, gibberellin metabolism, ethylene metabolism, jasmonic acid metabolism, transcription factors, R-genes, plant proteases, protein kinases, antifungal proteins, nitrogen transporters, sugar transporters, shikimate metabolism, isoflavone metabolism, phenylpropanoid metabolism, isoprenoid metabolism, â-oxidation lipid metabolism, fatty acid metabolism, glycolysis metabolism, gluconeogenesis metabolism, sucrose metabolism, sucrose catabolism, reductive pentose phosphate cycle, regulation of C3 photosynthesis, C4 pathway carbon assimilation, enzymes involved in the C4 pathway, carotenoid metabolism, tocopherol metabolism, phytosterol metabolism, brassinoid metabolism, and proline metabolism.

Site directed mutagenesis may be utilized to modify nucleic acid sequences, particularly as it is a technique that allows one or more of the amino acids encoded by a nucleic acid molecule to be altered (*e.g.*, a threonine to be replaced by a methionine) (Wells *et al.*, *Gene 34*:315-323 (1985); Gilliam *et al.*, *Gene 12*:129-137 (1980); Zoller and Smith, *Methods Enzymol. 100*:468-500 (1983); Dalbadie-McFarland *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 79*:6409-6413 (1982); Scharf *et al.*, *Science 233*:1076-1078 (1986); Higuchi *et al.*, *Nucleic Acids Res. 16*:7351-7367 (1988); U.S. Patent 5,811,238, European Patent 0 385 962; European Patent 0 359 472; and PCT Patent Application WO 93/07278; Lanz *et al.*, *J. Biol. Chem. 266*:9971-9976 (1991); Kovgan and Zhdanov, *Biotekhnologiya 5*:148-154, No. 207160n, Chemical Abstracts 110:225 (1989); Ge *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 86*:4037-4041 (1989); Zhu *et al.*, *J. Biol. Chem. 271*:18494-18498 (1996); Chu *et al.*, *Biochemistry 33*:6150-6157 (1994); Small *et al.*, *EMBO J. 11*:1291-1296 (1992); Cho *et al.*, *Mol. Biotechnol. 8*:13-16 (1997); Kita *et al.*, *J. Biol. Chem. 271*:26529-26535 (1996); Jin *et al.*, *Mol. Microbiol. 7*:555-562 (1993); Hatfield and Vierstra, *J. Biol. Chem. 267*:14799-14803 (1992); Zhao *et al.*, *Biochemistry 31*:5093-5099 (1992)).

Any of the nucleic acid molecules of the invention may either be modified by site directed mutagenesis or used as, for example, nucleic acid molecules that are used to target other nucleic acid molecules for modification.

It is understood that mutants with more than one altered nucleotide can be constructed using techniques that practitioners are familiar with, such as isolating restriction fragments and ligating such fragments into an expression vector (*see,* for example, Sambrook *et al., Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Press (1989)).

Two steps may be employed to characterize DNA-protein interactions. The first is to identify sequence fragments that interact with DNA-binding proteins, to titrate binding activity, to determine the specificity of binding and to determine whether a given DNA-binding activity can interact with related DNA sequences (Sambrook *et al., Molecular Cloning: A Laboratory Manual*, 2nd edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (1989)). Electrophoretic mobility-shift assay is a widely used assay. The assay provides a rapid and sensitive method for detecting DNA-binding proteins based on the observation that the mobility of a DNA fragment through a nondenaturing, low-ionic strength polyacrylamide gel is retarded upon association with a DNA-binding protein (Fried and Crother, *Nucleic Acids Res. 9*:6505-6525 (1981)). When one or more specific binding activities have been identified, the exact sequence of the DNA bound by the protein may be determined.

Several procedures for characterizing protein/DNA-binding sites are used (Maxam and Gilbert, *Methods Enzymol. 65*:499-560 (1980); Wissman and Hillen, *Methods Enzymol. 208*:365-379 (1991); Galas and Schmitz, *Nucleic Acids Res. 5*:3157-3170 (1978); Sigman *et al., Methods Enzymol. 208*:414-433 (1991); Dixon *et al., Methods Enzymol. 208*:414-433 (1991)). It is understood that one or more of the nucleic acid molecules of the invention may be utilized to identify a protein or fragment thereof that specifically binds to a nucleic acid molecule of the invention. It is also understood that one or more of the protein molecules or fragments thereof of the invention may be utilized to identify a nucleic acid molecule that specifically binds to it.

A two-hybrid system is based on the fact that proteins, such as transcription factors that interact (physically) with one another carry out many cellular functions. Two-hybrid systems have been used to probe the function of new proteins (Chien *et al., Proc. Natl. Acad. Sci. (U.S.A.) 88*:9578-9582 (1991); Durfee *et al., Genes Dev. 7*:555-569 (1993); Choi *et al., Cell 78*:499-512 (1994); Kranz *et al., Genes Dev. 8*:313-327 (1994)).

Interaction mating techniques have facilitated a number of two-hybrid studies of protein-protein interaction. Interaction mating has been used to examine interactions between small sets

of tens of proteins (Finley and Brent, *Proc. Natl. Acad. Sci. (U.S.A.) 91*:12098-12984 (1994)), larger sets of hundreds of proteins (Bendixen *et al., Nucl. Acids Res. 22*:1778-1779 (1994)) and to comprehensively map proteins encoded by a small genome (Bartel *et al., Nature Genetics 12*:72-77 (1996)). This technique utilizes proteins fused to the DNA-binding domain and proteins fused to the activation domain. They are expressed in two different haploid yeast strains of opposite mating type and the strains are mated to determine if the two proteins interact. Mating occurs when haploid yeast strains come into contact and result in the fusion of the two haploids into a diploid yeast strain. An interaction can be determined by the activation of a two-hybrid reporter gene in the diploid strain.

It is understood that the protein-protein interactions of protein or fragments thereof of the invention may be investigated using the two-hybrid system and that any of the nucleic acid molecules of the invention that encode such proteins or fragments thereof may be used to transform yeast in the two-hybrid system.

Computer Readable Media

The nucleotide sequence provided in SEQ ID NO: 1 through SEQ ID NO: 463,173 or fragment thereof, or complement thereof, or a nucleotide sequence at least 90% identical, preferably 95%, identical even more preferably 99% or 100% identical to the sequence provided in SEQ ID NO: 1 through SEQ ID NO: 463,173 or fragment thereof, or complement thereof, can be "provided" in a variety of mediums to facilitate use. Such a medium can also provide a subset thereof in a form that allows a skilled artisan to examine the sequences.

In a preferred embodiment of the present invention computer readable media may be prepared that comprise nucleic acid sequences where preferably at least 10%, preferably at least 25%, more preferably at least 50% and even more preferably at least 75%, 80%, 85%, 90% or 95% of the nucleic acid sequences are selected from the group of nucleic acid molecules that specifically hybridize to one or more nucleic acid molecule having a nucleic acid sequence selected from the group of SEQ ID NO: 1 through SEQ ID NO: 463,173 or complement thereof or fragments of either.

In another preferred embodiment of the present invention computer readable media may be prepared that comprise nucleic acid sequences where preferably at least 10%, preferably at least 25%, more preferably at least 50% and even more preferably at least 75%, 80%, 85%, 90%

or 95% of the nucleic acid sequences are selected from the group of nucleic acid molecules having a nucleic acid sequence selected from the group of SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof.

In a more preferred embodiment of the present invention, the computer readable media comprises a nucleic acid sequence and/or collection of nucleic acid sequences of the present invention associated with a biochemical process or activity where the process or activity is preferably selected from photosynthetic activity, carbohydrate metabolism, amino acid synthesis or degradation, plant hormone or other regulatory molecules, phenolic metabolism, and lipid metabolism, and more preferably selected from the group consisting of biosynthesis of tetrapyrroles, phytochrome metabolism, carbon assimilation, glycolysis and gluconeogenesis metabolism, sucrose metabolism, starch metabolism, phosphogluconate metabolism, galactomannan metabolism, raffinose metabolism, complex carbohydrate synthesis/degradation, phytic acid metabolism, methionine biosynthesis, methionine degradation, lysine metabolism, arginine metabolism, proline metabolism, glutamate/glutamine metabolism, aspartate/asparagine metabolism, cytokinin metabolism, gibberellin metabolism, ethylene metabolism, jasmonic acid metabolism, transcription factors, R-genes, plant proteases, protein kinases, antifungal proteins, nitrogen transporters, sugar transporters, shikimate metabolism, isoflavone metabolism, phenylpropanoid metabolism, isoprenoid metabolism, α-oxidation lipid metabolism, and fatty acid metabolism, and even more preferably selected from the group consisting of: glycolysis metabolism, gluconeogenesis metabolism, sucrose metabolism, sucrose catabolism, reductive pentose phosphate cycle, regulation of C3 photosynthesis, C4 pathway carbon assimilation, enzymes involved in the C4 pathway, carotenoid metabolism, tocopherol metabolism, phytosterol metabolism, brassinoid metabolism, and proline metabolism.

In an even more preferred embodiment of the present invention, the computer readable media comprises a nucleic acid sequence and/or collection of nucleic acid sequences of the present invention where the nucleic acid sequence and/or collection of nucleic acid sequences are associated with a component or attribute of at least two, more preferable at least three, four, five, six, seven, eight, nine, ten, eleven, twelve, thirteen, fourteen, fifteen, sixteen, seventeen, eighteen, nineteen, twenty, twenty one, twenty two, twenty three, twenty four, twenty five, twenty six, twenty seven, twenty eight, twenty nine, thirty, thirty one, thirty two, thirty three,

thirty four, thirty five, thirty six, thirty seven, thirty eight, thirty nine, forty, forty one, forty two, forty three, forty four, forty five or forty six biochemical processes or activities where the biochemical processes or activities are selected from the following: photosynthetic activity, carbohydrate metabolism, amino acid synthesis or degradation, plant hormone or other regulatory molecules, phenolic metabolism, lipid metabolism, biosynthesis of tetrapyrroles, phytochrome metabolism, carbon assimilation, glycolysis and gluconeogenesis metabolism, sucrose metabolism, starch metabolism, phosphogluconate metabolism, galactomannan metabolism, raffinose metabolism, complex carbohydrate synthesis/degradation, phytic acid metabolism, methionine biosynthesis, methionine degradation, lysine metabolism, arginine metabolism, proline metabolism, glutamate/glutamine, aspartate/asparagine metabolism, cytokinin metabolism, gibberellin metabolism, ethylene metabolism, jasmonic acid synthesis metabolism, transcription factors, R-genes, plant proteases, protein kinases, antifungal proteins, nitrogen transporters, sugar transporters, shikimate metabolism, isoflavone metabolism, phenylpropanoid metabolism, isoprenoid metabolism, β-oxidation lipid metabolism, fatty acid metabolism, glycolysis metabolism, gluconeogenesis metabolism, sucrose metabolism, sucrose catabolism, reductive pentose phosphate cycle, regulation of C3 photosynthesis, C4 pathway carbon assimilation, enzymes involved in the C4 pathway, carotenoid metabolism, tocopherol metabolism, phytosterol metabolism, brassinoid metabolism, and proline metabolism.

In one application of this embodiment, a nucleotide sequence of the present invention can be recorded on computer readable media. As used herein, "computer readable media" refers to any medium that can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as floppy discs, hard disc, storage medium and magnetic tape: optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media. A skilled artisan can readily appreciate how any of the presently known computer readable mediums can be used to create a manufacture comprising computer readable medium having recorded thereon a nucleotide sequence of the present invention.

As used herein, "recorded" refers to a process for storing information on computer readable medium. A skilled artisan can readily adopt any of the presently known methods for recording information on computer readable medium to generate media comprising the

nucleotide sequence information of the present invention. A variety of data storage structures are available to a skilled artisan for creating a computer readable medium having recorded thereon a nucleotide sequence of the present invention. The choice of the data storage structure will generally be based on the means chosen to access the stored information. In addition, a variety of data processor programs and formats can be used to store the nucleotide sequence information of the present invention on computer readable medium. The sequence information can be represented in a word processing text file, formatted in commercially-available software such as WordPerfect and Microsoft Word, or represented in the form of an ASCII file, stored in a database application, such as DB2, Sybase, Oracle, or the like. A skilled artisan can readily adapt any number of data processor structuring formats (*e.g.,* text file or database) in order to obtain computer readable medium having recorded thereon the nucleotide sequence information of the present invention.

By providing one or more of nucleotide sequences of the present invention, a skilled artisan can routinely access the sequence information for a variety of purposes. Computer software is publicly available which allows a skilled artisan to access sequence information provided in a computer readable medium. The examples which follow demonstrate how software which implements the BLAST (Altschul *et al., J. Mol. Biol. 215:*403-410 (1990)) and BLAZE (Brutlag *et al., Comp. Chem. 17:*203-207 (1993)) search algorithms on a Sybase system can be used to identify open reading frames (ORFs) within the genome that contain homology to ORFs or proteins from other organisms. Such ORFs are protein-encoding fragments within the sequences of the present invention and are useful in producing commercially important proteins such as enzymes used in amino acid biosynthesis, metabolism, transcription, translation, RNA processing, nucleic acid and a protein degradation, protein modification and DNA replication, restriction, modification, recombination and repair.

The present invention further provides systems, particularly computer-based systems, which contain the sequence information described herein. Such systems are designed to identify commercially important fragments of the nucleic acid molecule of the present invention. As used herein, "a computer-based system" refers to the hardware means, software means and data storage means used to analyze the nucleotide sequence information of the present invention. The minimum hardware means of the computer-based systems of the present invention comprises a

central processing unit (CPU), input means, output means and data storage means. A skilled artisan can readily appreciate that any one of the currently available computer-based system are suitable for use in the present invention.

As indicated above, the computer-based systems of the present invention comprise a data storage means having stored therein a nucleotide sequence of the present invention and the necessary hardware means and software means for supporting and implementing a search means. As used herein, "data storage means" refers to memory that can store nucleotide sequence information of the present invention, or a memory access means which can access manufactures having recorded thereon the nucleotide sequence information of the present invention. As used herein, "search means" refers to one or more programs that are implemented on the computer-based system to compare a target sequence or target structural motif with the sequence information stored within the data storage means. Search means are used to identify fragments or regions of the sequence of the present invention that match a particular target sequence or target motif. A variety of known algorithms are disclosed publicly and a variety of commercially available software for conducting search means are available can be used in the computer-based systems of the present invention. Examples of such software include, but are not limited to, MacPattern (EMBL), BLASTIN and BLASTIX (NCBIA). One of the available algorithms or implementing software packages for conducting homology searches can be adapted for use in the present computer-based systems.

The most preferred sequence length of a target sequence is from about 10 to 100 amino acids or from about 30 to 300 nucleotide residues. However, it is well recognized that during searches for commercially important fragments of the nucleic acid molecules of the present invention, such as sequence fragments involved in gene expression and protein processing, may be of shorter length.

As used herein, "a target structural motif," or "target motif," refers to any rationally selected sequence or combination of sequences in which the sequences the sequence(s) are chosen based on a three-dimensional configuration which is formed upon the folding of the target motif. There are a variety of target motifs known in the art. Protein target motifs include, but are not limited to, enzymatic active sites and signal sequences. Nucleic acid target motifs

include, but are not limited to, promoter sequences, *cis* elements, hairpin structures and inducible expression elements (protein binding sequences).

Thus, the present invention further provides an input means for receiving a target sequence, a data storage means for storing the target sequences of the present invention sequence identified using a search means as described above and an output means for outputting the identified homologous sequences. A variety of structural formats for the input and output means can be used to input and output information in the computer-based systems of the present invention. A preferred format for an output means ranks fragments of the sequence of the present invention by varying degrees of homology to the target sequence or target motif. Such presentation provides a skilled artisan with a ranking of sequences which contain various amounts of the target sequence or target motif and identifies the degree of homology contained in the identified fragment.

A variety of comparing means can be used to compare a target sequence or target motif with the data storage means to identify sequence fragments sequence of the present invention. For example, implementing software that implements the BLAST and BLAZE algorithms (Altschul *et al., J. Mol. Biol. 215:*403-410 (1990)) can be used to identify open frames within the nucleic acid molecules of the present invention. A skilled artisan can readily recognize that any one of the publicly available homology search programs can be used as the search means for the computer-based systems of the present invention.

Having now generally described the invention, the same will be more readily understood through reference to the following examples which are provided by way of illustration and are not intended to be limiting of the present invention, unless specified.

Example 1

This example illustrates the generation of libraries from cDNA prepared from a variety of *Arabidopsis thaliana*, Columbia ecotype, Landsberg ecotype, Nossen ecotype, *Glycine max, Zea mays* L., *Gossympium hirsutum, Sorghum bicolor, Oryza sativa* L (japonica type), *Oryza sativa* L (japonica type), *cv.* Nipponbare, *Zea mays* L. ssp *mexicana* and *Triticum aestivum* tissue. A subset of *Arabidopsis* libraries is used as an example.

Wild type *Arabidopsis thaliana* seeds are planted in commonly used planting pots and grown in an environmental chamber. Tissue is harvested as follows:

(a) For leaf tissue-based cDNA, leaf blades are cut with sharp scissors at seven weeks after planting;

(b) For root tissue-based cDNA, roots of seven-week old plants are rinsed intensively with tap water to wash away dirt, and briefly blotted by paper towel to take away free water;

(c) For stem tissue-based cDNA, stems are collected seven to eight weeks after planting by cutting the stems from the base and cutting the top of the plant to remove the floral tissue;

(d) For flower bud tissue-based cDNA, green and unopened flower buds are harvested about seven weeks after planting;

(e) For open flower tissue-based cDNA, completely opened flowers with all parts of floral structure observable, but no siliques are appearing, and are harvested about seven weeks after planting;

(f) For immature seed tissue-based cDNA, seeds are harvested at approximately 7-8 weeks of age. The seeds range in maturity from the smallest seeds that could be dissected from siliques to just before starting to turn yellow in color.

All tissue is immediately frozen in liquid nitrogen and stored at -80°C until total RNA extraction. The stored RNA is purified using Trizol reagent from Life Technologies (Gibco BRL, Life Technologies, Gaithersburg, Maryland U.S.A.), essentially as recommended by the manufacturer. Poly A+ RNA (mRNA) is purified using magnetic oligo dT beads essentially as recommended by the manufacturer (Dynabeads, Dynal Corporation, Lake Success, New York U.S.A.).

Construction of plant cDNA libraries is well known in the art and a number of cloning strategies exist. A number of cDNA library construction kits are commercially available. The Superscript™ Plasmid System for cDNA synthesis and Plasmid Cloning (Gibco BRL, Life Technologies, Gaithersburg, Maryland U.S.A.) is used, following the conditions suggested by the manufacturer.

The cDNA libraries are plated on LB agar containing the appropriate antibiotics for selection and incubated at 37° for a sufficient time to allow the growth of individual colonies.

Single selective media colonies are individually placed in each well of a 96-well microtiter plates containing LB liquid including the selective antibiotics. The plates are incubated overnight at approximately 37°C with gentle shaking to promote growth of the cultures. The plasmid DNA is isolated from each clone using Qiaprep plasmid isolation kits, using the conditions recommended by the manufacturer (Qiagen Inc., Santa Clara, California U.S.A.).

The template plasmid DNA clones are used for subsequent sequencing. For sequencing the cDNA libraries, a commercially available sequencing kit, such as the ABI PRISM dRhodamine Terminator Cycle Sequencing Ready Reaction Kit with AmpliTaq® DNA Polymerase, FS, is used under the conditions recommended by the manufacturer (PE Applied Biosystems, Foster City, CA). The cDNAs of the present invention are generated by sequencing initiated from the 5' end or 3' end of each cDNA clone. Entire inserts or only part of the inserts (ESTs or expressed sequenced tags) are sequenced.

A number of sequencing techniques are known in the art, including fluorescence-based sequencing methodologies. These methods have the detection, automation and instrumentation capability necessary for the analysis of large volumes of sequence data. Currently, the 377 DNA Sequencer (Perkin-Elmer Corp., Applied Biosystems Div., Foster City, CA) allows the most rapid electrophoresis and data collection. With these types of automated systems, fluorescent dye-labeled sequence reaction products are detected and data entered directly into the computer, producing a chromatogram that is subsequently viewed, stored, and analyzed using the corresponding software programs. These methods are known to those of skill in the art and have been described and reviewed (Birren *et al., Genome Analysis: Analyzing DNA*,1, Cold Spring Harbor, New York, the entirety of which is herein incorporated by reference).

The generated ESTs (including any full-length cDNA inserts or complete coding sequences) are combined with ESTs and full length cDNA sequences in public databases such as GenBank. Duplicate sequences are removed; and , duplicate sequence identification numbers are replaced. The combined dataset is then clustered and assembled using Pangea Systems tool identified as CAT v.3.2. First, the EST sequences are screened and filtered, *e.g.* high frequency words are masked to prevent spurious clustering; sequence common to known contaminants such as cloning bacteria are masked; high frequency repeated sequences and simple sequences are masked; unmasked sequences of less than 100 bp are eliminated. The thus-screened and filtered

ESTs are combined and subjected to a word-based clustering algorithm which calculates sequence pair distances based on word frequencies and uses a single linkage method to group like sequences into clusters of more than one sequence, as appropriate. Clustered sequence files are assembled individually using an iterative method based on PHRAP/CRAW/MAP providing one or more self-consistent consensus sequences and inconsistent singleton sequences. The assembled clustered sequence files are checked for completeness and parsed to create data representing each consensus contiguous sequence (contig), the initial EST sequences, and the relative position of each EST in a respective contig. The sequence of the 5' most clone is identified from each contig. The initial sequences that are not included in a contig are separated out. A FASTA file is created consisting of 463,173 sequences comprising the sequence of each contig and all original sequences which were not included in a contig. The EST contigs and original sequences which are not included in a contig are presented in Table 1 comprising SEQ ID No: 1 through SEQ ID NO: 463,173.

Example 2

The GenBank database is searched with BLASTN, version 2.0 (BLASTN takes a nucleotide sequence (the query sequence) and its reverse complement and searches them against a nucleotide sequence database) and BLASTX version 2.0 (BLASTX takes a nucleotide sequence, translates it in three forward reading frames and three reverse complement reading frames, and then compares the six translations against a protein sequence database) using default values with the cDNAs as queries. cDNA nucleic acid molecules that pass the E value threshold of $10e^{-8}$ for the following enzymes are classified. Results from these searches are set forth in Table 1.

References

Each reference mentioned in this specification is incorporated by reference in its entirety. In addition, these references, as well as each of those cited can be relied upon to make and use aspects of the invention.

Table 1

The entries in the Seq No. column refer to the corresponding sequence in the sequence listing.

Contig ID

The Contig ID is the name of the contig sequence found in the Monsanto SeqDB database.

Seq ID

The Seq ID heading is used to refer to a singleton. Seq ID is the name of the sequence found in the Monsanto Seq DB and also the name of the corresponding clone.

5' -most EST

Each contig is comprised of ESTs and/or full length insert sequences. The name of the most 5' sequence is listed in this column. The name of the sequence and the name of the most 5' clone are interchangable.

Method

The BLAST program used to determine homology, either BLASTX or BLASTN version 2.0 , is listed in this column.

NCBI GI

Each sequence in the GenBank public database is arbitrarily assigned a unique NCBI gi (National Center for Biotechnology Information GenBank Identifier) number. In this table, the NCBI gi number which is associated (in the same row) with a given Contig ID refers to the particular GenBank sequence which is used in the sequence comparison.

Blast Score

Bit score for BLAST match score that is generated by the sequence comparison of the cDNA with the GenBank sequence listed in the Description column. The E-value corresponding

to a given bit score is $E = mn2^{-S'}$. "m" and "n" are two proteins of length "m" and "n", "E" is the E value and S' is the bit score.

E-Value

The expectation E (range 0 to infinity) calculated for an alignment between the query sequence and a database sequence can be extrapolated to an expectation over the entire database search, by converting the pairwise expectation to a probability (range 0-1) and multiplying the result by the ratio of the entire database size (expressed in residues) to the length of the matching database sequence. In detail:

$$E\_database = (1 - \exp(-E))\, D / d$$

where D is the size of the database; d is the length of the matching database sequence; and the quantity (1 - exp(-E)) is the probability, P, corresponding to the expectation E for the pairwise sequence comparison.

Match Length

The match length is the number of identical residues between the query and subject sequences.

%Identity

The entries in the "%Ident" column of the table refer to the percentage of identically matched nucleotides (or residues) that exist along the length of that portion of the sequences which is aligned by the BLAST 2.0 comparison to generate the statistical scores presented.

% Similarity

Between the two lines of a Query and Subject (database) sequence in BLASTX output is a line indicating the specific residues which are identical, as well as those which are non-identical but nevertheless have positive alignment scores defined in the scoring matrix that is used.

The number of positives and the number of identities out of the total number possible is the % similarity score.

NCBI description

The "NCBI Description" column provides a description of the NCBI gi referenced in the "NCBIgi" column.

We claim:

1. A substantially purified nucleic acid molecule comprising a nucleic sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 463,173 or complements thereof or fragments of either.

2. A substantially purified first nucleic acid molecule, wherein the first nucleic molecule specifically hybridizes to a second nucleic acid molecule having a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 463,173 complements thereof.

3. A substantially purified protein or fragment thereof encoded by a first nucleic acid molecule which specifically hybridizes to a second nucleic acid molecule, the second nucleic acid molecule having a nucleic acid sequence selected from the group consisting of a complement of SEQ ID NO: 1 through SEQ ID NO: 463,173.

4. A substantially purified protein or fragment thereof encoded by a first nucleic acid molecule according to claim 3, wherein said first nucleic acid molecule comprises a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 463,173.

5. A purified antibody or fragment thereof which is capable of specifically binding to a protein or fragment thereof, wherein the protein or fragment thereof is encoded by a nucleic acid molecule comprising a nucleic acid sequence selected from the group consisting SEQ ID NO: 1 through SEQ ID NO: 463,173.

6. A transformed plant having a nucleic acid molecule which comprises: (A) an exogenous promoter region which functions in a plant cell to cause the production of a mRNA molecule; (B) a structural nucleic acid molecule comprising a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 463,173; and (C) a 3' non-translated sequence that functions in the plant cell to cause termination of transcription and addition of polyadenylated ribonucleotides to a 3' end of the mRNA molecule.

7. A transformed plant according to claim 6, wherein said plant is selected from the group consisting of maize, soybean, rice, wheat, or cotton.

8. A transformed plant having a nucleic acid molecule which comprises: (A) an exogenous promoter region which functions in a plant cell to cause the production of a mRNA molecule; which is linked to (B) a transcribed nucleic acid molecule with a transcribed strand

and a non-transcribed strand, wherein the transcribed strand is complementary to a nucleic acid molecule comprising a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 463,173 or fragment thereof; which is linked to (C) a 3' non-translated sequence that functions in plant cells to cause termination of transcription and addition of polyadenylated ribonucleotides to a 3' end of the mRNA molecule.

9. A transformed plant according to claim 8, wherein said plant is selected from the group consisting of maize, soybean, rice, wheat, or cotton.

10. A method for determining a level or pattern of a protein in a plant cell or plant tissue under evaluation which comprises assaying the concentration of a molecule, whose concentration is dependent upon the expression of a gene, the gene specifically hybridizes to a nucleic acid molecule having a nucleic acid sequence selected from the group consisting of a complement of SEQ ID NO: 1 through SEQ ID NO: 463,173, in comparison to the concentration of that molecule present in a reference plant cell or a reference plant tissue with a known level or pattern of the protein, wherein the assayed concentration of the molecule is compared to the assayed concentration of the molecule in the reference plant cell or reference plant tissue with the known level or pattern of the protein.

Abstract

The present invention is in the field of plant biochemistry. More specifically the invention relates to nucleic acid sequences from plant cells, in particular, nucleic acid sequences from maize, teosinte, soybean, *Arabidopsis*, cotton, Sorghum, rice and wheat. The invention encompasses nucleic acid molecules that encode proteins and fragments of proteins. In addition, the invention also encompasses proteins and fragments of proteins so encoded and antibodies capable of binding these proteins or fragments. The invention also relates to methods of using the nucleic acid molecules, proteins and fragments of proteins, and antibodies, for example for genome mapping, gene identification and analysis, plant breeding, preparation of constructs for use in plant gene expression, and transgenic plants.

United States Patent & Trademark Office

Office of Initial Patent Examination -- Scanning Division

SCANNED, # 15

Application deficiencies found during scanning:

□ Page(s)__________ of ____________________________ were not present for scanning.
(Document title)

□ Page(s)__________ of ____________________________ were not present for scanning.
(Document title)

* THE CDR REPLACES THE PAPER COPY OF SEQUENCE LISTINGS...

□ ***Scanned copy is best available.***